IOWA STATE UNIVERSITY
Digital Repository

Retrospective Theses and Dissertations

Iowa State University Capstones, Theses and Dissertations

1996

Spatio-temporal statistical models with application to atmospheric processes

Christopher Kim Wikle
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

 Part of the Atmospheric Sciences Commons

Recommended Citation
Wikle, Christopher Kim, "Spatio-temporal statistical models with application to atmospheric processes " (1996). *Retrospective Theses and Dissertations*. 11502.
https://lib.dr.iastate.edu/rtd/11502

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Spatio-temporal statistical models

with application to atmospheric processes

by

Christopher Kim Wikle

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Departments: Statistics;

Geological and Atmospheric Sciences

Majors: Statistics;

Meteorology

Major Professors: Noel Cressie and Tsing-Chang Chen

Iowa State University

Ames, Iowa

1996

Copyright © Christopher Kim Wikle, 1996. All rights reserved.

UMI Number: 9635368

UMI Microform 9635368
Copyright 1996, by UMI Company. All rights reserved.

This microform edition is protected against unauthorized
copying under Title 17, United States Code.

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of

Christopher K. Wikle

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Co-major Department

Signature was redacted for privacy.

For the Co-major Department

Signature was redacted for privacy.

For the Graduate College

TABLE OF CONTENTS

GENERAL INTRODUCTION

1 Background

Perhaps, it could be argued, we humans have a fundamental need to understand the world in which we live. At least, it is safe to say that it is usually in our best interest to do so. One feature of our world where this is certainly true is the atmosphere, specifically its long-term (i.e., climate) and short-term (i.e., weather) variability. In fact, one doesn't have to look too far to see the impact of weather and climate in virtually every aspect of our daily lives. Atmospheric influences are evident in agriculture, commerce, travel, recreation, and so forth. Consequently, characterization of atmospheric variability is more than just a curiosity, it is essential.

Broadly, we might say that the atmosphere/ocean system can be described as the superposition of a set of deterministic, multivariate, and nonlinear interactions over an enormous range of spatial and temporal scales. In order to understand this system, we must observe, summarize, make inference, and ultimately predict its behavior at each scale of variability, as well as the interaction between these scales. Unfortunately, although the system is deterministic in principle, our knowledge is incomplete at each of the observation, summarization, and inference stages, and thus our understanding of the atmosphere is clouded by uncertainty. Consequently, by the time we get to the prediction phase, our lack of certainty, combined with the nonlinear dynamics of the system, contributes to what is now known as *dynamical chaos*. As originally outlined by Lorenz(1963), chaos implies a fundamental lack of predictability. However, all is not lost, as over the last 100 years or so, the science of statistics has given us

many tools with which to evaluate, quantify, and exploit probabilistic uncertainty. Although we are always faced with the inherent chaotic nature of the atmosphere/ocean system, we can approach many of the relevant scientific questions from a probabilistic viewpoint, which allows us to make useful inferences in the presence of uncertainty, at least for relatively large spatial scales and relatively short temporal scales. Furthermore, we are then able to look for possible associations within and between variables in the system, which may allow us to extend our still incomplete physical theory.

Central to the observation, summarization, inference, and prediction of the atmosphere/ocean system is *data*. Unfortunately, all data come bundled with error. This is an inescapable fact of scientific life. In particular, along with the obvious errors associated with the measuring, manipulating, and archiving of data, there are errors due to the discrete spatial and temporal sampling of an inherently continuous system. Consequently, there are always scales of variability that are unresolvable, and which will surely contaminate the observations. In atmospheric science, this is considered a form of "turbulence", and corresponds to the well-known aliasing problem in time-series analysis (e.g., Chatfield 1989, p. 126) and the "nugget effect" in geostatistics (e.g., Cressie 1993, p. 59). Furthermore, atmospheric and oceanic data are rarely sampled at spatial or temporal locations that are optimal for the solution of a specific scientific problem. For instance, there is an obvious bias in data coverage towards areas where population density is large and, due to the cost of obtaining observations, towards a country whose Gross Domestic Product (GDP) is relatively large. Thus, the location of a measuring site and its temporal sampling frequency may have very little to do with science. Therefore, to gain scientific insight, we must consider these uncertainties when framing our questions, choosing our analysis techniques, and interpreting our results. This task is complicated further since atmospheric and oceanic data are nearly always correlated in space and time. In this case most of the traditional statistical methods taught in introductory statistical methods courses (which assume independent and identically distributed data) do not apply.

Traditionally, researchers in the atmospheric and related sciences have generally focused on

relatively simple time-series approaches and, outside of the data assimilation area of speciality, simple descriptive spatial techniques. These approaches have proven very valuable. However, as our already enormous data sets keep growing due to new observing platforms (e.g., satellite, radar, lidar, and profiler data), and as we ask more penetrating scientific questions with possibly severe implications (e.g., "given the observations to date, is there evidence of anthropogenic climate change?"), we must have more sophisticated techniques with which to handle the uncertainty in the data. Atmospheric scientists have risen to the challenge in recent years, and have exploited spatio-temporal methods such as Empirical Orthogonal Functions or EOFs (e.g., Lorenz 1956; Preisendorfer 1988), spatio-temporal Canonical Correlation Analysis or CCA (e.g., Glahn 1968; Bretherton et al. 1992), and Principal Oscillation Patterns or POPs (e.g., Hasselmann 1988; von Storch et al. 1988,1994). These are excellent tools with which to summarize data and have, to a lesser extent, sometimes been considered for prediction. However, often the assumptions under which these methods should be used are either poorly understood or ignored. Furthermore, as typically used, such methods contain no mechanism for spatial prediction. This is unfortunate since incomplete spatial sampling of the atmosphere makes spatial prediction an issue of critical importance. It is apparent that there is a need for additional spatio-temporal methods in the atmospheric sciences.

There have been some notable recent attempts to systematically introduce statistical ideas to the atmospheric science community (e.g., Thiébaux 1994; Wilks 1995), although the scope of these works is relatively broad. My goal in this dissertation is to continue to narrow the gap between the need for atmospheric science related spatio-temporal methods, and modern statistical approaches to such methods. To do this, I will first provide an overview of spatio-temporal statistical methods commonly used in the atmospheric sciences, from a statistical point of view. Next, I will use statistical techniques ranging from the very simple (e.g., harmonic analysis) to "state-of-the-art" (e.g., autoregressive cyclic spectral analysis) to characterize the variability of outstanding atmospheric science problems concerned with the spatial structure of the extratropical semi-annual cycle, and temporal variability of mixed-Rossby gravity waves over the

tropical Pacific. Finally, I will develop and implement a new and very general spatio-temporal statistical model that is both spatially descriptive and temporally dynamic. This model is applied to monthly precipitation data over the South China Sea.

2 Dissertation Organization

The dissertation is organized according to a "paper format". Following the General Introduction, a preliminary chapter is included to provide an overview of spatio-temporal statistical methods in the atmospheric sciences. This is then followed by three studies which have been, or will be, submitted for publication in an appropriate atmospheric science or statistics journal. Each paper is self-contained and includes a full literature review. Since each paper is independent, equation numbers only apply to the paper at hand. The papers are then followed by a general conclusion. A brief summary of each chapter is described below.

2.1 Spatio-Temporal Statistical Methods in the Atmospheric Sciences

In this overview, I examine the traditional spatio-temporal statistical methods used in the atmospheric sciences. These methods are considered from a statistical perspective. Although this section is primarily a review, many of the statistical issues that are considered have not, to my knowledge, been considered in the context of these methods. As a consequence, several "open questions" are posed in this review.

2.2 On the Semiannual Variation in the Northern Hemisphere Extratropical Height Field

Based upon the application to atmospheric data of a recent signal-processing technique for identifying periodic components in the presence of unknown noise (Wikle et al. 1995), it became clear that there is a distinct semiannual oscillation (SAO) in most atmospheric variables. It was also clear that the strength of this SAO signal varies depending on spatial location. Since no unified view of the extratropical Northern Hemisphere (NH) SAO has been published, I

sought to find a mechanism by which the SAO spatial variation could be characterized. Along with Professor T.-C. Chen, I discovered that the NH midlatitude SAO in 500hPa geopotential height could be explained almost entirely as a result of spatial and temporal asymmetries in the annual variation of the stationary eddies. We then concluded that the mechanism for the SAO in the NH is simply a result of land-sea contrasts, similar to the mechanism proposed by van Loon (1967) for the Southern Hemisphere (SH) SAO. There is a fundamental difference, however, in that the NH extratropics are dominated by *east-west* land-sea contrasts due to the large continental land masses in the NH, while the SH land-sea contrast reflects the *north-south* differential heating between Antarctica and the surrounding ocean. These results are important primarily for their application to sensitivity tests of atmospheric general circulation models (GCMs). In particular, before we can accept GCM results related to interannual and interdecadal variability, we must be sure that they can properly simulate the relatively simple components of the atmospheric seasonal cycle, namely the SAO. To date, they are *not* able to do this.

2.3 Seasonal Variation of Lower Stratospheric Mixed Rossby-Gravity Waves over the Tropical Pacific

It has long been understood that the atmospheric general circulation is principally driven by diabatic heating with its centers located in the tropics. This diabatic heating is mostly attributable to the latent heat released from cumulus convection. Moreover, it has been shown that tropical planetary waves can be modulated by this diabatic heating. Thus, since tropical convective activity is known to show semiannual variability (e.g., Chen and Wu 1992), we would expect to see a semiannual signal in equatorial wave activity.

Along with Dr. R. Madden of the National Center for Atmospheric Research (NCAR) and Professor T.-C. Chen, I examined the seasonal variability of mixed Rossby-gravity waves (MRGWs) in lower stratospheric wind measurements over the equatorial Pacific. Utilizing the seasonally varying spectral analysis procedure developed by Madden (1986) and the recently developed autoregressive cyclic spectral analysis technique (e.g., Sherman and White 1995), I

found that there are significant twice-yearly peaks in MRGW activity, with peaks occurring in winter-early spring and in summer-early fall. In addition, the seasonally varying spectral analysis suggested that there is a previously unknown convergence of horizontal momentum flux associated with these waves, and that the sign of that convergence is different during the times of the two seasonal maxima. There was also an indication from the cyclic spectral analysis that the frequency of the MRGWs may be different during the two maxima.

2.3.1 A Spatially Descriptive, Temporally Dynamic Statistical Model with Applications to Atmospheric Processes

In this paper, with Professor N. Cressie, a new spatio-temporal statistical model is proposed that attempts to consider the influence of both temporal and spatial variability. Unlike the traditional spatio-temporal methods used in the atmospheric science literature (and outlined in the first chapter of this dissertation), this method is mainly concerned with *prediction* in space and time. Our predictive model is temporally dynamic in that it exploits the unidirectional flow of time in an autoregressive framework. In addition, the model is spatially descriptive in the sense that although spatial correlation is modeled by a spatially colored noise process, no causative interpretation is associated with this noise. With the inclusion of measurement error, this formulation naturally leads to the development of a spatio-temporal Kalman filter. We can then use this Kalman filter to predict at future times and at locations for which we do not have data. We demonstrate this method through simulation and show its utility to atmospheric science by using it to predict monthly precipitation throughout the data-sparse South China Sea region.

References

Bretherton, C.S., C. Smith, and J.M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, **5**, 541-560.

Chatfield, C., 1989: *The Analysis of Time Series: An Introduction*, Fourth Edition, Chapman and Hall, 241pp.

Chen, T.-C., and K.D. Wu, 1992: Semi-annual oscillation of the global divergent circulation. *Tellus,* **44A**, 357-365.

Cressie, N.A.C, 1993: *Statistics for Spatial Data, Revised Edition*, Wiley, 900pp.

Glahn, H.R., 1968: Canonical correlation and its relationship to discriminant analysis and multiple regression. *J. Atmos. Sci.*, **25**,23-31.

Hasselmann, K., 1988: PIPs and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns. *J. Geophys. Res.*,**93**, 11015 - 11021.

Lorenz, E.N., 1956: Empirical orthogonal functions and statistical weather prediction. *Sci. Rept. No. 1, Statistical Forecasting Project*, MIT, 49 pp.

Lorenz, E.N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*,**20**,130-141.

Madden, R.A., 1986: Seasonal variations of the 40-50 day oscillation in the tropics. *J. Atmos. Sci.*,**43**, 3138-3158.

Preisendorfer, R.W., 1988: *Principal Component Analysis in Meteorology and Oceanography*, Elsevier, 425 pp.

Sherman, P.J., and L.B. White, 1995: Improved periodic spectral analysis with application to diesel vibration data. *J. Acoust. Soc. Amer.,* **98**, 3285-3301.

Thiébaux, H.J., 1994: *Statistical Data Analysis for Ocean and Atmospheric Sciences*, Academic Press, 247pp.

van Loon, H., 1967: The half-yearly oscillation in middle and high southern latitudes and the coreless winter. *J. Atmos. Sci.,* **24,** 472-486.

von Storch, H., G. Bürger, R. Schnur, and J.-S. von Storch, 1995: Principal oscillation patterns: a review. *J. Climate,* **8**, 377-400.

von Storch, H., T. Bruns, I. Fischer-Bruns, and K. Hasselmann, 1988: Principal oscillation pattern analysis of the 30- to 60-day oscillation in a general circulation model equatorial troposphere. *J. Geophys. Res.,* **93**, 11022-11036.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*, Academic Press, 467pp.

Wikle, C.K., P.J. Sherman, and T.-C. Chen, 1995: Identifying periodic components in atmospheric data using a family of minimum variance spectral estimators. *J. Climate,* **8**, 2352-2363.

SPATIO-TEMPORAL STATISTICAL METHODS IN THE ATMOSPHERIC SCIENCES

1 Introduction

Since virtually all meteorological and climatological processes involve variability over both space and time, it is imperative that statistical models for these processes also consider spatio-temporal variability. The atmospheric science literature contains many examples of statistical methods that capture various forms of spatio-temporal variability, both in diagnostic and prognostic applications. These methods include Empirical Orthogonal Function (EOF), Principal Interaction Pattern (PIP), Principal Oscillation Pattern (POP), and Canonical Correlation Analysis (CCA) techniques. There are often underlying assumptions and approximations associated with the application of these techniques that do not seem to be recognized or, at least, they are not reported. In particular, issues such as discrete vs. continuous space and time, prognostic vs. diagnostic application, measurement error vs. no measurement error, and the statistical criterion against which the methodology under consideration is optimal are not always considered explicitly.

My goal in this review is to examine the traditional spatio-temporal statistical methods used in the atmospheric sciences from a more statistical perspective. In addition, I shall try to emphasize the practical scientific motivation for these methods as they are (or should be) applied in the atmospheric sciences. Many of the statistical issues that are considered have not, to my knowledge, been considered in the literature. Thus, there may be several "open questions" that are posed along the way.

1.1 Notation and General Model Assumptions

Assume that we are given observations of a spatio-temporal process $Z(\mathbf{s};t)$ at spatial locations $\mathbf{s} \in \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ in some spatial domain D_z and time $t \in \{1, 2, \ldots, T\}$, where D_z is assumed to be two-dimensional Euclidean space unless otherwise noted. We may also have observations of some process $X(\mathbf{r};t)$ at spatial locations $\mathbf{r} \in \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m\} \subset D_x$ and time $t \in \{1, 2, \ldots, T\}$, where D_x is some spatial domain such that D_z and D_x may overlap, depending on the application. In some cases we will assume that there is measurement error associated with the observation of these processes.

2 Empirical Orthogonal Function (EOF) Analysis

EOF analysis is the geophysicist's manifestation of the classic eigenvalue/eigenvector decomposition of a correlation (or covariance) matrix. In its discrete formulation, EOF analysis is simply Principal Component Analysis or PCA (e.g., Hotelling 1933), while in the continuous framework, it is simply a Karhunen-Loève (K-L) expansion (e.g., Loève 1963). Depending on the application, EOFs are usually used:

- in a diagnostic mode to find principal (in terms of explanation of variance) spatial structures, along with the corresponding time variation of these structures, and
- to reduce the degrees of freedom (spatially) in large geophysical data sets while simultaneously reducing noise.

An extensive review of EOFs, their theory, and their application to meteorology and oceanography is contained in Preisendorfer (1988).

One finds in the meteorological literature, extensive use of EOFs since their introduction by Lorenz (1956). For example, they have been used for describing climate (e.g., Kutzbach 1967; Barnett 1977), for comparing simulations of general circulation models (e.g., Preisendorfer and Barnett 1983), for developing regression forecast techniques (e.g., Peagle and Haslam 1982), in weather classification (e.g., Christensen and Bryson 1966), in map typing (e.g., Richman 1981),

in the interpretation of geophysical fields (Olbed and Creutin 1986), and in the simulation of random fields, particularly non-homogeneous processes (Braud and Obled 1991). In addition, as in the psychometric literature for PCAs and as advocated by Richman (1981,1986) and others, orthogonal and oblique rotation of EOFs often aids in the interpretation of meteorological data. Furthermore, because EOFs have difficulty resolving traveling wave disturbances, complex EOF analysis was introduced by Wallace and Dickinson (1972) and was shown to be very useful in applications to certain meteorological problems (e.g., Wallace 1972; Barnett 1983; Horel 1984).

In the remainder of this section, I will briefly present the K-L expansion, show how the discrete EOF formulation can be derived in a PCA context, and present a couple of issues related to EOFs and measurement error that have not been considered in the literature. Finally, I will briefly consider the central idea behind some variations of EOF analysis, namely, complex EOF analysis, multivariate EOF analysis, and extended EOF analysis.

2.1 Continuous K-L Formulation

We first consider a continuous spatial process measured at discrete time intervals. Our goal is to find an optimal and separable orthogonal decomposition of a spatio-temporal process $Z(\mathbf{s};t)$. That is, we want

$$Z(\mathbf{s};t) = \sum_{k=1}^{\infty} a_k(t)\phi_k(\mathbf{s}) \tag{1}$$

such that the $\text{var}(a_1(t)) > \text{var}(a_2(t)) > \ldots$, and $\text{cov}(a_i(t), a_k(t)) = 0$ for all $i \neq k$. A well-known solution to this problem is obtained through a Karhunen-Loève (K-L) expansion (e.g., see Papoulis 1965, p. 457-461). Suppose,

$$\text{E}[Z(\mathbf{s};t)] = 0, \tag{2}$$

and define the covariance function as

$$\text{E}[Z(\mathbf{s};t)Z(\mathbf{r};t)] \equiv c_0^Z(\mathbf{s},\mathbf{r}), \tag{3}$$

which need not be stationary in space, but is assumed to be invariant in time. The K-L expansion then allows the covariance function to be decomposed as follows:

$$c_0^Z(\mathbf{s},\mathbf{r}) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\mathbf{s})\phi_k(\mathbf{r}), \tag{4}$$

where $\{\phi_k(\cdot) : k = 1, \ldots, \infty\}$ are the eigenfunctions and $\{\lambda_k : k = 1, \ldots, \infty\}$ are the associated eigenvalues of the Fredholm integral equation

$$\int_D c_0^Z(\mathbf{s},\mathbf{r})\phi_k(\mathbf{s})\, d\mathbf{s} = \lambda_k \phi_k(\mathbf{r}), \tag{5}$$

where

$$\int_D \phi_k(\mathbf{s})\phi_l(\mathbf{s}) d\mathbf{s} = \begin{cases} 1 & \text{for } k = l, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Assuming completeness of the eigenfunctions, we can then expand $Z(\mathbf{s};t)$ according to

$$Z(\mathbf{s};t) = \sum_{k=1}^{\infty} a_k(t)\phi_k(\mathbf{s}), \tag{7}$$

where we call $\{\phi_k(\mathbf{s}) : \mathbf{s} \in D\}$ the k-th EOF and often refer to the associated time series $a_k(t)$ as the k-th principal component time series, or "amplitude" time series. This time series is derived from the projection of the Z process onto the EOF basis,

$$a_k(t) = \int_D Z(\mathbf{s};t)\phi_k(\mathbf{s})\, d\mathbf{s}. \tag{8}$$

It is easy to verify that these time series are uncorrelated, with variance equal to the corresponding eigenvalues; that is,

$$\mathrm{E}[a_i(t)a_k(t)] = \delta_{ik}\lambda_k, \tag{9}$$

where δ_{ik} equals one when $i = k$, and equals zero otherwise.

If we truncate the expansion (7) at K, yielding

$$Z_K(\mathbf{s};t) \equiv \sum_{k=1}^{K} a_k(t)\phi_k(\mathbf{s}), \tag{10}$$

then it can be shown (e.g., Freiberger and Grenander 1965; Davis 1976) that the finite EOF decomposition minimizes the variance of the truncation error, $\mathrm{E}\{[Z(\mathbf{s};t) - Z_K(\mathbf{s};t)]^2\}$, and is thus optimal in this regard when compared to all other basis sets.

Since data are always discrete, in practice we must solve numerically the Fredholm integral equation (5) to obtain the EOF basis functions. Cohen and Jones (1986) and Buell (1972,1975) give numerical quadrature solutions to this problem. The numerical quadrature approaches for discretizing the integral equation succeed in that they give estimates for the eigenfunctions and eigenvalues that are weighted according to the spatial distribution of the data locations , but only for the eigenfunctions at locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ for which there are data. Obled and Creutin (1986) have presented an elegant method for discretizing the K-L integral equation and for interpolating the eigenfunctions to locations where data are not available. See the third paper of this dissertation for an example of this approach.

2.2 Discrete EOF Analysis

Although the continuous K-L representation of EOFs is the most realistic from a physical point-of-view, it is only rarely considered in applications. This is due simply to the discrete nature of data observations and the added difficulty of solving the K-L integral equation. Let us consider a discrete EOF analysis by using the PCA formulation as given in standard multivariate statistics books (e.g., Johnson and Wichern 1992), but according to the spatio-temporal notation we have introduced. In that case, let

$$\mathbf{Z}(t) \equiv (Z(\mathbf{s}_1; t), \ldots, Z(\mathbf{s}_n; t))' \tag{11}$$

and define the k-th EOF ($k = 1, \ldots, n$) to be

$$\boldsymbol{\psi}_k \equiv (\psi_k(\mathbf{s}_1), \ldots, \psi_k(\mathbf{s}_n))', \tag{12}$$

where $\boldsymbol{\psi}_k$ is the vector in the linear combination

$$a_k(t) = \boldsymbol{\psi}_k' \mathbf{Z}(t). \tag{13}$$

Furthermore, $\boldsymbol{\psi}_1$ is the vector that allows $\text{var}(a_1(t))$ to be maximized subject to the constraint $\boldsymbol{\psi}_1' \boldsymbol{\psi}_1 = 1$. Then $\boldsymbol{\psi}_2$ is the vector that maximizes $\text{var}(a_2(t))$ subject to the constraint $\boldsymbol{\psi}_2' \boldsymbol{\psi}_2 = 1$ and $\text{cov}(a_1(t), a_2(t)) = 0$. Thus, $\boldsymbol{\psi}_k$ is the vector that maximizes $\text{var}(a_k(t))$ subject to the

orthogonality constraint

$$\boldsymbol{\psi}_k'\boldsymbol{\psi}_k = 1 \tag{14}$$

and

$$\mathrm{cov}(a_k(t), a_j(t)) = 0, \text{ for all } k \neq j. \tag{15}$$

This is equivalent to solving the eigensystem

$$\mathbf{C}_0^Z \boldsymbol{\Psi} = \boldsymbol{\Psi}\boldsymbol{\Lambda}, \tag{16}$$

where

$$\mathbf{C}_0^Z \equiv \mathrm{E}[\mathbf{Z}(t)\mathbf{Z}(t)'] \tag{17}$$

$$\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n) \tag{18}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \tag{19}$$

and where

$$\mathrm{var}(a_i(t)) = \lambda_i, \quad i = 1, \ldots, n. \tag{20}$$

Then the solution of (1.16) is obtained by a symmetric decomposition of $\mathbf{C}_0^Z$,

$$\mathbf{C}_0^Z = \boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}', \tag{21}$$

which is the PCA formulation.

It is straightforward to show (e.g., Cohen and Jones 1969; Buell 1972) that if a discretization of the K-L integral equation assumes equal areas of influence for each observation location, then such a discretization is equivalent to the PCA formulation of EOFs. Conversely, an EOF decomposition of irregularly spaced data without consideration of the relative area associated with each observation location leads to improper weighting of the significance of each element of the covariance matrix $\mathbf{C}_0^Z$. This can give erroneous results in the EOF analysis. The distinction between EOFs on a regular grid and on an irregular grid is the source of many incorrect applications of the technique in the literature. For a discussion of the effects of ignoring this distinction, see Karl et al. (1982).

2.3 EOFs in the Presence of Measurement Error

The EOF decomposition in the presence of measurement error has not been considered in the literature. We assume that the process $Z(\cdot\,;\cdot)$ is the sum of a smooth signal of interest Y and additive noise,

$$\mathbf{Z}(t) = \mathbf{Y}(t) + \boldsymbol{\epsilon}(t), \tag{22}$$

where

$$\mathbf{Y}(t) \equiv (Y(\mathbf{s}_1;t),\ldots,Y(\mathbf{s}_n;t))' \tag{23}$$

$$\boldsymbol{\epsilon}(t) \equiv (\epsilon(\mathbf{s}_1;t),\ldots,\epsilon(\mathbf{s}_n;t))'. \tag{24}$$

Taking the variance of $\mathbf{Z}(t)$ gives

$$\mathbf{C}_0^Z = \mathbf{C}_0^Y + \mathbf{C}_0^\epsilon, \tag{25}$$

where

$$\mathbf{C}_0^Y \equiv \text{var}[\mathbf{Y}(t)] \tag{26}$$

$$\mathbf{C}_0^\epsilon \equiv \text{var}[\boldsymbol{\epsilon}(t)]. \tag{27}$$

Now, assuming regularly spaced observations, we can show that

$$\boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}' = \mathbf{C}_0^Y + \mathbf{C}_0^\epsilon. \tag{28}$$

If we let the measurement error be white noise with variance σ_ϵ^2, then

$$\mathbf{C}_0^\epsilon = \sigma_\epsilon^2\mathbf{I} = \sigma_\epsilon^2\boldsymbol{\Psi}\boldsymbol{\Psi}', \tag{29}$$

and

$$\mathbf{C}_0^Y = \boldsymbol{\Psi}(\boldsymbol{\Lambda} - \sigma_\epsilon^2\mathbf{I})\boldsymbol{\Psi}'. \tag{30}$$

Now, if we let

$$Y(\mathbf{s};t) = \sum_{k=1}^{n} a_k^*(t)\psi_k(\mathbf{s}), \tag{31}$$

where

$$a_k^*(t) \equiv \boldsymbol{\psi}_k'\mathbf{Y}(t). \tag{32}$$

Then,

$$\mathrm{E}[\mathbf{a}^*(t)\mathbf{a}^*(t)'] = \mathbf{\Lambda} - \sigma_\epsilon^2 \mathbf{I}, \tag{33}$$

where

$$\mathbf{a}^*(t) \equiv (a_1^*(t), \ldots, a_n^*(t))'. \tag{34}$$

Thus, the noisy Z process and the smooth Y process have the same eigenfunctions $\{\psi_k\}$, but different eigenvalues (i.e., $\{\lambda_k\}$ and $\{\lambda_k - \sigma_\epsilon^2\}$, respectively). In addition, the amplitude time series for both processes have the same mean (zero), but the variance of the Y process amplitude time series amplitude is less (as expected). Thus, if white measurement error is present (and it always is with observational data), then the spatial EOF spatial patterns (i.e., the $\{\psi_k\}$) are not affected, but the variance explained by each pattern must be adjusted accordingly. This is seldom (if ever) done in practice.

2.4 EOFs as a Method of Smoothing

The truncated EOF expansion in the discrete framework is given by

$$Z_K(\mathbf{s};t) = \sum_{k=1}^{K} a_k(t)\psi_k(\mathbf{s}). \tag{35}$$

As was the case for the EOF truncation in continuous space (10), the EOF basis can be shown to minimize the variance of the truncation error (e.g., Davis 1976). In practice, it is common to assume that this truncation can be used as a prediction of the smooth process Y. That is,

$$\hat{Y}(\mathbf{s};t) \equiv Z_K(\mathbf{s};t). \tag{36}$$

Although not considered in the literature, it is natural to ask if this is an optimal predictor.

To investigate the optimality of (36), we consider the space-only process analogous to (22), that is,

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}). \tag{37}$$

Assume we have data at n locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and that we are interested in predicting $Y(\mathbf{s}_i)$ from $\mathbf{Z} \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$. If we assume that $Z(\cdot)$ is Gaussian, then the optimal predictor

is just a linear predictor. After Cressie (1993, p. 110), this optimal predictor is given by

$$\mathrm{E}[Y(\mathbf{s}_i) \mid \mathbf{Z}] = \mathbf{c}^Y(\mathbf{s}_i)'[\mathbf{C}^Z]^{-1}\mathbf{Z}\,; \quad i = 1, \ldots, n\,, \tag{38}$$

where we have assumed that the means of $\mathbf{Z}$ and $Y(\mathbf{s}_i)$ are known to be zero, and we define

$$\mathbf{c}^Y(\mathbf{s}_i) \equiv \mathrm{cov}(Y(\mathbf{s}_i), \mathbf{Y}) \tag{39}$$

$$\mathbf{C}^Z \equiv \mathrm{var}(\mathbf{Z}) \tag{40}$$

and

$$\mathbf{Y} \equiv (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'. \tag{41}$$

Now, if we let $\mathrm{E}[\epsilon(\mathbf{s}_i)\epsilon(\mathbf{s}_k)] = \sigma_\epsilon^2$ for $i = k$, and equal to zero otherwise, then we can write

$$\mathbf{C}^Z = \mathbf{C}^Y + \sigma_\epsilon^2\mathbf{I}, \tag{42}$$

where

$$\mathbf{C} \equiv \mathrm{var}(\mathbf{Y}). \tag{43}$$

Now, since $\mathbf{C}^Y$ is real and symmetric, we can write the following decomposition:

$$\mathbf{C}^Y = \mathbf{\Psi}\mathbf{\Lambda}^*\mathbf{\Psi}', \tag{44}$$

where the eigenvectors and eigenvalues are:

$$\mathbf{\Psi} \equiv (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n) \tag{45}$$

$$\boldsymbol{\psi}_k \equiv (\psi_k(\mathbf{s}_1), \ldots, \psi_k(\mathbf{s}_n))', \quad k = 1, \ldots, n \tag{46}$$

$$\mathbf{\Lambda}^* \equiv \mathrm{diag}(\lambda_1^*, \ldots, \lambda_n^*)\,. \tag{47}$$

Then, using the orthogonality $\mathbf{\Psi}\mathbf{\Psi}' = \mathbf{\Psi}'\mathbf{\Psi} = \mathbf{I}$, we obtain

$$\mathbf{C}^Z = \mathbf{\Psi}\mathbf{\Lambda}^*\mathbf{\Psi}' + \sigma_\epsilon^2\mathbf{\Psi}\mathbf{\Psi}' \tag{48}$$

$$= \mathbf{\Psi}\mathbf{\Lambda}\mathbf{\Psi}', \tag{49}$$

where

$$\mathbf{\Lambda} = \mathbf{\Lambda}^* + \sigma_\epsilon^2\mathbf{I}. \tag{50}$$

Now, (38) can be written

$$\mathrm{E}[Y(\mathbf{s}_i) \mid \mathbf{Z}] = \mathbf{c}^Y(\mathbf{s}_i)'\mathbf{\Psi}\mathbf{\Lambda}^{-1}\mathbf{\Psi}'\mathbf{Z}. \tag{51}$$

Furthermore, since

$$\mathrm{E}[Y(\mathbf{s}_i)Y(\mathbf{s}_j)] = \sum_{k=1}^{n} \lambda_k^* \psi_k(\mathbf{s}_i)\psi_k(\mathbf{s}_j), \tag{52}$$

then

$$\mathbf{c}^Y(\mathbf{s}_i) = \mathbf{\Psi}\mathbf{\Lambda}^*\boldsymbol{\psi}(\mathbf{s}_i), \tag{53}$$

where

$$\boldsymbol{\psi}(\mathbf{s}_i) \equiv (\psi_1(\mathbf{s}_i), \ldots, \psi_n(\mathbf{s}_i))'; \quad i = 1, \ldots, n \ . \tag{54}$$

In addition, as for the EOF case,

$$\mathbf{Z} = \mathbf{\Psi}\mathbf{a}, \tag{55}$$

where

$$\mathbf{a} \equiv (a_1, \ldots, a_n)', \tag{56}$$

and

$$a_k = \boldsymbol{\psi}_k'\mathbf{Z}. \tag{57}$$

We can then write the optimal predictor as

$$\begin{aligned} \mathrm{E}[Y(\mathbf{s}_i) \mid \mathbf{Z}] &= \boldsymbol{\psi}(\mathbf{s}_i)\mathbf{\Lambda}^*\mathbf{\Psi}'\mathbf{\Psi}\mathbf{\Lambda}^{-1}\mathbf{\Psi}'\mathbf{\Psi}\mathbf{a} && (58) \\ &= \sum_{k=1}^{n} \frac{\lambda_k^*}{\lambda_k^* + \sigma_\epsilon^2} a_k \psi_k(\mathbf{s}_i). && (59) \end{aligned}$$

Now, the truncated predictor of interest is given by

$$Z_K(\mathbf{s}_i) = \sum_{k=1}^{K} a_k \psi_k(\mathbf{s}_i) \ ; i = 1, \ldots, n. \tag{60}$$

Thus, the optimal predictor (59) is equivalent to the truncated predictor (60) in the trivial case where $\sigma_\epsilon^2 = 0$ and $\sum_{l=K+1}^{n} a_l\psi_l(\mathbf{s}_i) = 0$. In general, setting (59) equal to (60) gives

$$\sum_{k=1}^{K} \frac{\lambda_k^*}{\lambda_k^* + \sigma_\epsilon^2} a_k \psi_k(\mathbf{s}_i) + \sum_{l=k+1}^{n} \frac{\lambda_l^*}{\lambda_l^* + \sigma_\epsilon^2} a_l \psi_l(\mathbf{s}_i) = \sum_{k=1}^{K} a_k \psi_k(\mathbf{s}_i), \tag{61}$$

which results in the equality constraint

$$\sum_{k=1}^{K}\left(\frac{\lambda_k^*}{\lambda_k^*+\sigma_\epsilon^2}-1\right)a_k\psi_k(\mathbf{s}_i)=-\sum_{l=K+1}^{n}\frac{\lambda_l^*}{\lambda_l^*+\sigma_\epsilon^2}a_l\psi_l(\mathbf{s}_i)\ ;i=1,\ldots,n. \tag{62}$$

Clearly, (62) shows that if $\sigma_\epsilon^2 = 0$, then optimality is achieved if $a_l = 0$ or $\lambda_l^* = 0$ for $l = K+1,\ldots,n$. Thus, in the absence of measurement error, if the truncation parameter K is large so that the λ_l^* are very close to zero, then the truncated EOF smoothing technique is "nearly" optimal. However, in the presence of measurement error, the discrete orthogonality of the eigenvectors in (62) implies optimality when $a_k = 0$ for $k = 1,\ldots,K$ and $\lambda_l^* = 0$ or $a_l = 0$ for $l = K+1,\ldots,n$. Under such conditions, one would probably not be interested in an EOF analysis. Thus, we can state that $Z_K(\mathbf{s}_i)$ is *not in general* an optimal predictor of $Y(\mathbf{s}_i)$. Consequently, the smoothed quantity $Z_K(\mathbf{s}_i)$ should be viewed as an *ad hoc* method for removing noise.

2.5 Estimation of EOFs

Since the EOF analysis depends on the decomposition of a covariance matrix, we must estimate this matrix in practice. The traditional approach is based on the method of moments (MOM) estimation procedure. For example, in the discrete case with equally spaced observations, we need an estimate of

$$\mathbf{C}_0^Z = \mathrm{E}[\mathbf{Z}(t)\mathbf{Z}(t)'], \tag{63}$$

where $\mathbf{Z}(t)$ is assumed to have zero mean. The MOM estimator for an element of $\mathbf{C}_0^Z$,

$$c_0^Z(\mathbf{s}_i,\mathbf{s}_j)\equiv \mathrm{E}(Z(\mathbf{s}_i;t)Z(\mathbf{s}_j;t)),\ \forall t, \tag{64}$$

is given by

$$\hat{c}_0^Z(\mathbf{s}_i,\mathbf{s}_j)\equiv (1/T)\sum_{t=1}^{T}[Z(\mathbf{s}_i;t)-\hat{\mu}_Z(\mathbf{s}_i;t)][Z(\mathbf{s}_j;t)-\hat{\mu}_Z(\mathbf{s}_j;t)]\ , \tag{65}$$

where $\hat{\mu}_Z(\mathbf{s}_i;t)$ is an estimate of the mean of $Z(\mathbf{s}_i;t)$, for $i = 1,\ldots,n$. This mean correction must be included since data quite typically show a nonzero mean. Possible choices of $\hat{\mu}_Z(\mathbf{s}_i;t)$

include the time mean,

$$\hat{\mu}_Z(\mathbf{s}_i) \equiv (1/T)\sum_{t=1}^{T} Z(\mathbf{s}_i;t), \tag{66}$$

the space mean,

$$\hat{\mu}_Z(t) \equiv (1/n)\sum_{i=1}^{n} Z(\mathbf{s}_i;t), \tag{67}$$

and the grand mean,

$$\hat{\mu}_Z \equiv (1/nT)\sum_{t=1}^{T}\sum_{i=1}^{n} Z(\mathbf{s}_i;t). \tag{68}$$

To the best of my knowledge, the investigation of the proper choice for estimating the mean has not appeared in the literature. Typically, investigators use the time mean (66), but it is not at all clear that this is the best choice. Further investigation is needed. Perhaps, until a study of the effect of the different choices can be performed, the best thing is to use the grand mean (68).

Given an estimate $\hat{\mathbf{C}}_0^Z$ of $\mathbf{C}_0^Z$ that is symmetric and non-negative definite (so that all eigenvalues are greater than or equal to zero), an estimate of its eigenvectors and eigenvalues can be obtained through the diagonalization

$$\hat{\mathbf{C}}_0^Z = \hat{\mathbf{\Psi}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Psi}}'. \tag{69}$$

Note that Lawley (1956) derived approximate formulas for the bias and variance of the standard eigenvalue estimator. Later, von Storch and Hannoschock (1985) extended Lawley's results and considered the bias and variance in the EOF case. Based on theory and Monte Carlo simulation, they found that the sample eigenvalue $\hat{\lambda}_k$ is a biased estimator of λ_k. This bias is positive for the larger λ_k's and negative for the smaller λ_k's. They note that unbiased estimators can be constructed, but that the decrease in bias is accompanied by an increase in the variance of the estimator.

Furthermore, it has been shown that the sampling error associated with the estimated EOFs leads to numerical instability in the eigenvectors (e.g., Gray 1981; North et al. 1982). This has led to sampling-based selection strategies for the truncation level, K. Many of these are described in Preisendorfer (1988).

2.6 Variations on the Standard EOF Analysis

In this section, I shall briefly examine the idea behind several extensions of the standard EOF analysis described above.

2.6.1 Complex EOF Analysis

Consider a spatio-temporal process consisting of a sinusoid in one spatial dimension that is invariant in time:

$$Z(s;t) = B\sin(ls), \tag{70}$$

where s is some location in one-dimensional space, t is a time index, B is an amplitude coefficient, and l is the spatial wave number, which is related to the wavelength L such that $l = 2\pi/L$. Now, consider the same sinusoid but allow it to have a temporal phase component (i.e., it can be considered as a wave in space which propagates in time):

$$\begin{aligned} Z(s;t) &= B\sin(ls+\omega t) && (71)\\ &= B\cos(\omega t)\sin(ls) - B\sin(\omega t)\cos(ls), && (72) \end{aligned}$$

where ω is the temporal frequency. Thus, as the difference between (70) and (72) clearly shows, in order to characterize the phase propagation of such a sinusoid, we need information regarding the coefficients of the *two* components, $\sin(ls)$ and $\cos(ls)$, which are a quarter of a cycle out of phase. In time series analysis, this is analogous to the need for both the quadrature and co-spectrum between two time series in order to determine their spectral coherence and phase relationships (e.g., Chatfield 1989).

One advantage of the EOF approach described previously is its ability to compress the complicated variability of the original data set onto a relatively small set of eigenvectors. Unfortunately, such an EOF analysis only detects spatial structures that do not change position in time (analogous to equation (70) above). To extend the EOF analysis to the study of spatial structures that can propagate in time (e.g., analogous to equation (71) above), Wallace and Dickinson (1972) developed complex principal component analysis in the frequency domain.

The technique involves the computation of complex eigenvectors from cross-spectral matrices. The limitation of this technique is that it only gives the decomposition for individual (i.e., very narrow) frequency bands. Consequently, if the power of a phenomenon is spread over a wide frequency band (as is generally the case with physical phenomena), then several EOF spatial maps (one for each spectral estimate) are needed to evaluate the phenomenon. This complicates the physical interpretation.

Complex empirical orthogonal function (CEOF) analysis in the time domain was developed as an alternative to the frequency-domain approach described above. It was originally presented by Rasmusson et al. (1981), and has since been used by a number of investigators (e.g., Barnett 1983; Trenberth and Shin 1984). This method differs from the frequency-domain approach in that Hilbert transforms (see below) are used to shift the time series of the data at each location by a quarter cycle. Analogous to (72), the original data and its Hilbert transform allow the examination of propagating disturbances. Horel (1984) gives an excellent discussion of some of the theoretical and practical issues related to CEOF analysis. In our development, the method is described following the approach of Barnett (1983).

Consider $\{Z(\mathbf{s}_j;t)\ ;j=1,\ldots,n\}$ as described previously. Under certain regularity conditions, $Z(\mathbf{s}_j;t)$ has a Fourier representation of the form

$$Z(\mathbf{s}_j;t)=\sum_{\omega}\alpha_j(\omega)\cos(\omega t)+\beta_j(\omega)\sin(\omega t), \tag{73}$$

where $\alpha_j(\omega)$ and $\beta_j(\omega)$ are the Fourier coefficients, and ω is the frequency $(-\pi \leq \omega \leq \pi)$. Since the description of propagating features requires phase information, it is convienient to use the complex representation:

$$Z^f(\mathbf{s}_j;t)=\sum_{\omega}\gamma_j(\omega)e^{-i\omega t}, \tag{74}$$

where $\gamma_j(\omega)=\alpha_j(\omega)+i\beta_j(\omega)$. Using the definition of $\gamma_j(\omega)$, we can expand (74) to obtain:

$$Z^f(\mathbf{s}_j;t)=Z(\mathbf{s}_j;t)+i\tilde{Z}(\mathbf{s}_j;t), \tag{75}$$

where

$$Z(\mathbf{s}_j;t)=\alpha_j(\omega)\cos(\omega t)+\beta_j(\omega)\sin(\omega t), \tag{76}$$

and,

$$\tilde{Z}(\mathrm{s}_j;t) = \beta_j(\omega)\cos(\omega t) - \alpha_j(\omega)\sin(\omega t). \tag{77}$$

The real part $Z(\mathrm{s}_j;t)$ is the original process and the imaginary part $\tilde{Z}(\mathrm{s}_j;t)$ is the Hilbert transform of the original process, which is just the original process with its phase shifted in time by $\frac{\pi}{2}$. Barnett (1983) gives several methods for estimating the Hilbert transform.

Now, the covariance matrix of $Z^f(\mathrm{s}_j;t)$ can be written as:

$$\mathbf{C}_0^{Zf} = [c_0^{Zf}(\mathrm{s}_j,\mathrm{s}_k)]_{j,k=1,\ldots,n} \tag{78}$$

where

$$c_0^{Zf}(\mathrm{s}_j,\mathrm{s}_k) \equiv \mathrm{E}[Z^f(\mathrm{s}_j;t)^* Z^f(\mathrm{s}_k;t)], \tag{79}$$

and where $*$ denotes the complex conjugate. Note that $\mathbf{C}_0^{Zf}$ is essentially the cross-spectral matrix averaged over all frequencies ($-\pi \leq \omega \leq \pi$), and thus leads to an average depiction of the propagating disturbances present in the data. If we are only interested in phenomena occurring over a certain spectral frequency range of ω, then we can filter accordingly the original process $Z(\cdot\,;t)$ and its Hilbert transform $\tilde{Z}(\cdot\,;t)$ before the CEOF analysis.

Since $\mathbf{C}_0^{Zf}$ is Hermitian, it possesses real eigenvalues $\{\lambda_k\}$ and complex eigenvectors,

$$\gamma_k \equiv (\gamma_k(\mathrm{s}_1),\ldots,\gamma_k(\mathrm{s}_n))'\,; \quad k = 1,\ldots,n. \tag{80}$$

The EOF representation of $Z^f(\cdot\,;t)$, which optimally accounts for the variance of $Z(\cdot\,;t)$ in the frequency band of interest, is:

$$Z^f(\mathrm{s}_i;t) = \sum_{k=1}^{n} a_k(t)\gamma_k^*(\mathrm{s}_i), \tag{81}$$

where the complex time-dependent principal components are given by:

$$a_k(t) = \sum_{i=1}^{n} Z^f(\mathrm{s}_i;t)\gamma_k(\mathrm{s}_i). \tag{82}$$

Four measures are generally used to examine the structure of the CEOFs.

- *Spatial Phase Function.* The spatial phase function is given by:

$$\theta_k(\mathbf{s}_i) = \arctan\left[\frac{Im(\gamma_k(\mathbf{s}_i))}{Re(\gamma_k(\mathbf{s}_i))}\right]. \tag{83}$$

This function can take any value between $-\pi$ and π. In the case of the simple sinusoid with temporal phase (71), this corresponds to ls. In that case the spatial phase will go through one complete cycle (2π) over the distance $2\pi/l$. It should be noted that for data fields which include many different scales of variability, the spatial phase plot can be very difficult to interpret. The pre-filtering procedure described above generally improves interpretability.

- *Spatial Amplitude Function.* The spatial amplitude function is given by:

$$S_k(\mathbf{s}_i) = [\gamma_k(\mathbf{s}_i)\gamma_k^*(\mathbf{s}_i)]^{1/2}. \tag{84}$$

This function is interpreted in the same way as the eigenfunctions in traditional EOF analysis.

- *Temporal Phase Function.* The temporal phase function is given by:

$$\xi_k(t) = \arctan\left[\frac{Im(a_k(t))}{Re(a_k(t))}\right]. \tag{85}$$

Consider the simple sinusoid example in (71). For a fixed frequency ω_0 ,this temporal phase function would give $\omega_0 t$ (i.e., a linear relationship in time). In practice, this provides information as to the frequency of the dominant component of a particular eigenvector at a given time.

- *Temporal Amplitude Function.* The temporal amplitude function is given by:

$$R_k(t) = [a_k(t)a_k^*(t)]^{1/2}. \tag{86}$$

This function corresponds to the amplitude time series as given in traditional EOF analysis.

2.6.2 Multivariate EOF Analysis

Often, we may be interested in the simultaneous analysis of two or more processes. Kutzbach (1967) used a form of EOF analysis which simultaneously considered several meteorological variables at many spatial locations and times. Preisendorfer (1988) considers this approach at length. A brief description of the basic idea behind this multivariate EOF analysis methodology follows.

Consider two fields observed over time at the same spatial locations; that is, consider $Z(\mathbf{s}_i;t)$ and $X(\mathbf{s}_i;t)$, where $i = 1,\ldots,n$; $t = 1,\ldots,T$. Then, we can write

$$\mathbf{W}(t) \equiv [\mathbf{Z}(t)'\mathbf{X}(t)']', \tag{87}$$

where

$$\mathbf{X}(t) \equiv (X(\mathbf{s}_1;t),\ldots,X(\mathbf{s}_n;t))', \tag{88}$$

and $\mathbf{Z}(t)$ is defined in (11). Then, the covariance matrix of $\mathbf{W}(t)$ is given by

$$\mathbf{C}_0^W = \mathrm{E}[\mathbf{W}(t)\mathbf{W}(t)']. \tag{89}$$

Thus, it is clear that this matrix includes off-diagonal submatrices that represent the covariance between $\mathbf{Z}(t)$ and $\mathbf{X}(t)$. As shown in Preisendorfer (1988, p.161-162), one can then obtain the EOF solution in the conventional manner by diagonalizing the $\mathbf{C}_0^W$ matrix; that is,

$$\mathbf{C}_0^W = \mathbf{\Psi}_W \mathbf{\Lambda}_W \mathbf{\Psi}_W', \tag{90}$$

where the columns of $\mathbf{\Psi}_W$ are the eigenvectors (i.e., EOFs) and $\mathbf{\Lambda}_W$ is a diagonal matrix containing the eigenvalues of $\mathbf{C}_0^W$. Then, the first n elements of the k-th eigenvector correspond to the portion of the k-th EOF for the Z process, and the last n elements correspond to the portion representative of the k-th EOF of the X process. Theoretically, there is no limit to the number of processes that could be considered simultaneously. However, there is a practical limitation to this procedure since the covariance matrix (89) can easily become very large if the number of observation locations or variables increases. Bretherton et al. (1992) compare

this approach to other multivariate methods such as Canonical Correlation Analysis and Singular Value Decomposition (see Section 5) and find that, in some cases, the multivariate EOF approach has large biases and does not perform well in small signal-to-noise ratio situations.

2.6.3 Extended EOF Analysis

Extended EOFs (e.g., Weare and Nasstrom 1982) are simply multivariate EOFs in which the additional variables are lagged versions of the same process. For example, we could let

$$\mathbf{W}(t) = [\mathbf{Z}(t)'\mathbf{Z}(t-1)']' \tag{91}$$

In this case, if temporal invariance is assumed, then the diagonal sub-matrices of $\mathbf{C}_0^W$ are equivalent, and the off-diagonal submatrices are just the lag-one correlation matrices

$$\mathbf{C}_1^Z \equiv \mathrm{E}[\mathbf{Z}(t)\mathbf{Z}(t-1)']. \tag{92}$$

In this way, we can examine the propagation of EOF spatial patterns in time by noting that the first n eigencoefficients of a particular eigenvector correspond to the time zero representation of that EOF, and the second n eigencoefficients correspond to the lag one representation of the same EOF. This approach is closely linked with time-lagged CCA and the minimum/maximum autocorrelation factor (MAF) method in statistics. A brief comparison of these three approaches is presented in Section 5.3.

3 Principal Interaction Patterns (PIPs)

Principal Interaction Patterns (PIPs) were originally proposed by Hasselmann (1988) for the continuous time case. He considered a system represented by the state vector $\mathbf{Z}(t)$, whose evolution is governed by a set of first-order equations,

$$\frac{d\mathbf{Z}(t)}{dt} = \mathcal{F}(\mathbf{Z}(t)), \tag{93}$$

where $\mathcal{F}$ is some (possibly) non-linear time-dependent function of $\mathbf{Z}(t)$. The goal is to construct a simplified dynamical model approximating (93) which involves a significantly smaller number

of dimensions $m : m < n$. von Storch et al. (1995) have explained PIPs from the discrete-time perspective using vector-space ideas. We shall adopt the discrete-time approach as well, although from a generalized least-squares point of view.

Our goal is to approximate the dynamical system $\mathbf{Z}(t) \in R^n$ as being driven by a lower-dimensional dynamical system $\mathbf{a}(t) \in R^m$. There is a statistical model underlying PIPs, which we write in the following form

$$\mathbf{Z}(t) = \mathbf{P}\mathbf{a}(t) + \epsilon(t) \tag{94}$$

$$\mathbf{a}(t+1) = \mathcal{F}(\mathbf{a}(t), \boldsymbol{\beta}, t) + \boldsymbol{\eta}(t), \tag{95}$$

where $\mathbf{Z}(t)$ are the observations at locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and at time t; $\mathbf{P} \equiv [\mathbf{p}_1, \ldots, \mathbf{p}_m]$ is an $n \times m$ matrix with column vectors $\mathbf{p}_i$, known as the PIPs; $\mathcal{F}(\cdot)$ denotes a class of models that can be nonlinear in the dynamical variables $\mathbf{a}(t)$ and, additionally, depends on a set of "free" parameters $\boldsymbol{\beta}$; and $\epsilon(t), \boldsymbol{\eta}(t)$ are, in general, unspecified error terms. We shall assume here that $\epsilon(t)$ is multivariate white noise and is uncorrelated with $\boldsymbol{\eta}(t)$.

The PIP analysis is given below as a two-stage procedure.

Stage One:

Define the predictor

$$\hat{\mathbf{Z}}(t) \equiv \mathbf{P}\mathbf{a}(t). \tag{96}$$

Assuming $\mathbf{P}$ known and $\text{var}[\epsilon(t)] \equiv \mathbf{R}$, the generalized least squares estimator of $\mathbf{a}(t)$ can be obtained by minimizing

$$(\mathbf{Z}(t) - \mathbf{P}\mathbf{a}(t))'\mathbf{R}^{-1}(\mathbf{Z}(t) - \mathbf{P}\mathbf{a}(t)) \tag{97}$$

with respect to $\mathbf{a}(t)$. Then the estimator of $\mathbf{a}(t)$ is

$$\hat{\mathbf{a}}(t) = (\mathbf{P}'\mathbf{R}^{-1}\mathbf{P})^{-1}\mathbf{P}'\mathbf{R}^{-1}\mathbf{Z}(t). \tag{98}$$

Stage Two:

Now assume $\mathbf{a}(t)$ is known and seek estimates of $\mathbf{P}$ and $\boldsymbol{\beta}$ by a generalized least squares difference between the derivatives of the $\mathbf{Z}(\cdot)$ process and the predictor (96). Because we are

approaching the problem from a discrete time perspective, we consider the first-differences rather than the derivatives.

$$\begin{aligned} \mathbf{U}(t) &\equiv \mathbf{Z}(t+1) - \mathbf{Z}(t) && (99) \\ &= \mathbf{P}[\mathbf{a}(t+1) - \mathbf{a}(t)] + \boldsymbol{\epsilon}(t+1) - \boldsymbol{\epsilon}(t) && (100) \\ &= \mathbf{P}[\mathcal{F}(\mathbf{a}(t), \boldsymbol{\beta}, t) - \mathbf{a}(t)] + \boldsymbol{\nu}(t), && (101) \end{aligned}$$

where

$$\boldsymbol{\nu}(t) \equiv \mathbf{P}\boldsymbol{\eta}(t) + \boldsymbol{\epsilon}(t+1) - \boldsymbol{\epsilon}(t). \tag{102}$$

Define

$$\hat{\mathbf{U}} \equiv \mathbf{P}[\mathcal{F}(\mathbf{a}(t), \boldsymbol{\beta}, t) - \mathbf{a}(t)], \tag{103}$$

and

$$\text{var}[\boldsymbol{\nu}(t)] \equiv \mathbf{V}. \tag{104}$$

Then, a generalized least-squares estimator of $\mathbf{P}$ and $\boldsymbol{\beta}$ is obtained by minimizing

$$(\mathbf{U}(t) - \hat{\mathbf{U}}(t))'\mathbf{V}^{-1}(\mathbf{U}(t) - \hat{\mathbf{U}}(t)) \tag{105}$$

with respect to $\mathbf{P}$ and $\boldsymbol{\beta}$.

This minimization is complicated by the fact that $\mathbf{V}$ is a function of $\mathbf{P}$ as can be seen from (102). Furthermore, if $\mathcal{F}(\cdot)$ is nonlinear, then some additional estimation complexity is present. Either way, the minimization of (105) is likely to require iteration. The two-stage formulation suggests that (98) and (105) could be used in an iterative procedure.

3.1 Application of PIPs

Until recently, a full implementation of PIPs had not been performed in the atmospheric sciences. As is usually the case for nonlinear regression with additive errors, the nonlinear function $\mathcal{F}(\cdot)$ should have some physical justification. Realistic models in the atmospheric sciences are quite complex systems of nonlinear equations and this complexity has slowed the implementation of PIPs. However, Achatz et al. (1995) have used PIPs in the examination

of baroclinic wave life cycles. To the best of my knowledge, no one has tried a completely empirical approach to PIPs in which they let the data define the structure of the $\mathcal{F}(\cdot)$ function. One simplification of PIPs, which has been used extensively, is known as Principal Oscillation Patterns (POPs) and will be examined in Section 4.

3.2 Alternate Views of PIPs

We note that equations (94) and (95) could be implemented in a Kalman filter framework. In particular, since the function $\mathcal{F}(\cdot)$ is in general nonlinear, one would need to make use of one of the nonlinear Kalman filter approaches, such as the extended Kalman filter. This technique essentially uses the linear term in the Taylor series expansion of $\mathcal{F}(\cdot)$ (e.g., Grewal and Andrews 1993, p.168-170). Perhaps surprisingly, this approach has not been used in the PIP literature. However, one could argue that this is the approach that has effectively been tried in some of the recent experimental Kalman filter approaches to the data assimilation problem in meteorology (e.g., Miller et al. 1994; Daley 1995), although it has not been recognized as such.

Because the first differences (99) are used in the estimation of parameters, we might also consider developing the model in a multivariate context in terms of $\mathbf{Z}(t)$ and $\mathbf{U}(t)$, where

$$\mathbf{Z}(t) = \mathbf{P}\mathbf{a}(t) + \boldsymbol{\epsilon}(t) \tag{106}$$

$$\mathbf{U}(t) = \mathbf{P}[\mathcal{F}(\mathbf{a}(t), \boldsymbol{\beta}, t) - \mathbf{a}(t)] + \boldsymbol{\nu}(t) \tag{107}$$

$$\mathbf{a}(t+1) = \mathcal{F}(\mathbf{a}(t), \boldsymbol{\beta}, t) + \boldsymbol{\eta}(t), \tag{108}$$

and

$$\boldsymbol{\nu}(t) \equiv \mathbf{P}\boldsymbol{\eta}(t) + \boldsymbol{\epsilon}(t+1) - \boldsymbol{\epsilon}(t). \tag{109}$$

Such a system has not been investigated. However, it seems that it would be superior to the traditional PIP model because it includes cross-covariances between $\mathbf{Z}(t)$ and $\mathbf{U}(t)$, in a manner similar to that for cokriging in the geostatistical literature (e.g., Cressie 1993, p.138-142).

4 Principal Oscillation Patterns (POPs)

Principal Oscillation Patterns (POPs) were originally formulated as a specific case of PIPs by Hasselmann (1988). POPs were reformulated by von Storch et al. (1988) and extended to complex fields by Bürger (1993). Recent years have seen a substantial increase in the number of POP applications in the literature. A comprehensive overview of POPs can be found in von Storch et al. (1995).

In essence, POP analysis assumes that the (multivariate) data field has a temporal autoregressive structure of order one (AR(1)). A chief difference between POPs and other spatio-temporal decompositions (with the exception of PIPs) is that the eigenvectors (i.e., the spatial patterns) are not orthonormal. In addition, the eigenvectors can be characterized as the empirical "normal modes" of the (estimated) system matrix of the fitted AR(1) stochastic process.

4.1 Formulation of POPs

We hypothesize the following AR(1) model for the dynamical process $\mathbf{Z}(t)$:

$$\mathbf{Z}(t+1) = \mathbf{B}\mathbf{Z}(t) + \boldsymbol{\eta}(t), \tag{110}$$

where $\mathbf{B}$ is an $n \times n$ real matrix (possibly non-symmetric), $\boldsymbol{\eta}(t)$ is an $n \times 1$ additive error vector (often assumed to be Gaussian white noise) such that

$$\mathrm{E}[\boldsymbol{\eta}(t)\mathbf{Z}(t)'] = \mathbf{0} \tag{111}$$

$$\mathrm{E}[\boldsymbol{\eta}(t)] = \mathbf{0} \tag{112}$$

$$\mathrm{E}[\boldsymbol{\eta}(t)\boldsymbol{\eta}(\tau)'] \equiv \mathbf{C}^{\eta}, \ t = \tau; \ \mathbf{0}, \text{ otherwise.} \tag{113}$$

Thus, this model is a special case of the PIP model (94) and (95) in which there is no reduction in the order of the dynamical system (i.e., $\mathbf{Z}(t) = \mathbf{a}(t)$) and the two equations collapse into one. The function $\mathcal{F}(\cdot)$ is the matrix product using $\mathbf{B}$, and the system is linear and first-order Markov.

Note that if (110) is post-multiplied by $\mathbf{Z}(t)$ then, upon taking expectations, we obtain

$$\mathrm{E}[\mathbf{Z}(t+1)\mathbf{Z}(t)'] = \mathbf{B}\mathrm{E}[\mathbf{Z}(t)\mathbf{Z}(t)'] + \mathrm{E}[\boldsymbol{\eta}(t)\mathbf{Z}(t)']. \tag{114}$$

Thus, we can solve for $\mathbf{B}$ in

$$\mathbf{B} = \mathrm{E}[\mathbf{Z}(t+1)\mathbf{Z}(t)'](\mathrm{E}[\mathbf{Z}(t)\mathbf{Z}(t)'])^{-1} \quad (115)$$

$$= \mathbf{C}_1^Z[\mathbf{C}_0^Z]^{-1}, \quad (116)$$

where

$$\mathbf{C}_1^Z \equiv \mathrm{cov}[\mathbf{Z}(t+1), \mathbf{Z}(t)] \quad (117)$$

$$\mathbf{C}_0^Z \equiv \mathrm{var}[\mathbf{Z}(t)], \quad (118)$$

and we have assumed $\mathrm{E}[\mathbf{Z}(t)]$ and $\mathrm{E}[\mathbf{Z}(t+1)]$ to be zero, and that $\mathbf{C}_0^Z$ is non-singular.

As shown by Rao (1973, p.43-44), for a non-symmetric matrix $\mathbf{A}_{n\times n}$, the characteristic equation $|\mathbf{A} - \lambda\mathbf{I}| = \mathbf{0}$ has n roots, some of which may be complex even if $\mathbf{A}$ is real. Corresponding to a latent root λ_i there are two vectors $\mathbf{p}_i, \mathbf{q}_i$ called the "right" and "left" singular vectors, such that

$$\mathbf{A}\mathbf{p}_i = \lambda_i\mathbf{p}_i\,; \quad i = 1, \ldots, n \quad (119)$$

$$\mathbf{A}'\mathbf{q}_i = \lambda_i\mathbf{q}_i\,; \quad i = 1, \ldots, n\,. \quad (120)$$

Rao then shows that

- $\{\mathbf{p}_i\}$ are linearly independent and so are $\{\mathbf{q}_i\}$,
- $\mathbf{p}_i^*\mathbf{q}_j = 0$ for $i \neq j$, where * denotes the Hermitian transpose (e.g., $\mathbf{a}^*\mathbf{b} = \sum_i a_i\bar{b}_i$),
- If we let $\mathbf{p}_i^*\mathbf{q}_i \equiv d_i$ and $\mathbf{D} \equiv \mathrm{diag}(d_i)$: $i = 1, \ldots, n$, then $\mathbf{A}$ has the spectral decomposition

$$\mathbf{A} = \mathbf{P}\Lambda\mathbf{D}^{-1}\mathbf{Q}' \quad (121)$$

$$= \sum_{i=1}^{n} \frac{\lambda_i}{d_i}\mathbf{p}_i\mathbf{q}_i^*, \quad (122)$$

where we have assumed that the latent roots are distinct.

Using (122), we can decompose the AR(1) model matrix $\mathbf{B}$ in (110) as

$$\mathbf{B} = \sum_{i=1}^{n} \frac{\lambda_i}{d_i}\phi_i\psi_i^*, \quad (123)$$

where $d_i \equiv \psi_i^* \phi_i$, ϕ_i is the right singular vector of $\mathbf{B}$, and ψ_i is the left singular vector of $\mathbf{B}$, corresponding to the latent root $\lambda_i : \ i = 1, \ldots, n$.

Now, note that because

$$\sum_{i=1}^{n} \frac{\phi_i \psi_i^*}{d_i} = \mathbf{\Phi D}^{-1} \mathbf{\Psi}^* \tag{124}$$

$$= \mathbf{\Phi} (\mathbf{\Psi}^* \mathbf{\Phi})^{-1} \mathbf{\Psi}^* \tag{125}$$

$$= \mathbf{I}_{n \times n}, \tag{126}$$

we can write

$$\mathbf{Z}(t) = \sum_{i=1}^{n} \frac{\phi_i \psi_i^*}{d_i} \mathbf{Z}(t) \tag{127}$$

$$= \sum_{i=1}^{n} a_i(t) \phi_i \ , \tag{128}$$

where

$$a_i(t) \equiv \frac{\psi_i^* \mathbf{Z}(t)}{d_i}. \tag{129}$$

Thus, the essence of the POP analysis is the decomposition represented by (128) and (129). The vectors $\{\phi_i\}$ are called *principal oscillation patterns* and, although they constitute a linear basis, they are *not orthonormal.* The time series $a_i(t)$ are known as POP coefficients. We note that, although the ϕ_i are not orthogonal with themselves, they are orthogonal with the normalized adjoint patterns ψ_i^*/d_i.

4.2 Physical Implication of POPs

To gain insight into the physical meaning of the POPs, we consider an idealized discrete linear system (i.e., with no error term),

$$\mathbf{Z}(t+1) = \mathbf{B}\mathbf{Z}(t). \tag{130}$$

The eigenvectors ϕ_i; $i = 1, \ldots, n$ of the $\mathbf{B}$ matrix are referred to as the system *normal modes.* Because $\mathbf{B}$ is not in general symmetric, some or all of its eigenvalues and eigenvectors are

complex. Furthermore, because $\mathbf{B}$ is real, the complex conjugates λ_i^* and $\boldsymbol{\phi}_i^*$ also satisfy the eigen-equation:

$$\mathbf{B}\boldsymbol{\phi}_i^* = \lambda_i^*\boldsymbol{\phi}_i^*. \tag{131}$$

Now, if we multiply (130) by $\boldsymbol{\psi}_j^*/d_j$, we obtain

$$\begin{aligned} a_j(t+1) &= \frac{\boldsymbol{\psi}_j^*}{d_j}[\sum_{i=1}^{n}\frac{\lambda_i}{d_i}\boldsymbol{\phi}_i\boldsymbol{\psi}_i^*]\mathbf{Z}(t) && (132)\\ &= \frac{\boldsymbol{\psi}_j^*}{d_j}\sum_{i=1}^{n}\lambda_i\boldsymbol{\phi}_i a_i(t) && (133)\\ &= \frac{\boldsymbol{\psi}_j^*}{d_j}\boldsymbol{\phi}_j\lambda_j a_j(t) && (134)\\ &= \lambda_j a_j(t). && (135) \end{aligned}$$

Equation (135) is a first-order homogeneous difference equation with solution (assuming $a_j(0) = 1$)

$$a_j(t) = (\lambda_j)^t. \tag{136}$$

If λ_j is complex and $i \equiv \sqrt{-1}$, then

$$\lambda_j \equiv \lambda_j^R + i\lambda_j^I, \tag{137}$$

which can be written in polar form as

$$\begin{aligned} \lambda_j^R &= \gamma_j\cos(\omega_j) && (138)\\ \lambda_j^I &= \gamma_j\sin(\omega_j). && (139) \end{aligned}$$

Then,

$$\lambda_j = \gamma_j e^{i\omega_j}, \tag{140}$$

where

$$\gamma_j = \{(\lambda_j^R)^2 + (\lambda_j^I)^2\}^{1/2}. \tag{141}$$

Thus, (136) and (140) gives

$$a_j(t) = \gamma_j^t e^{i\omega_j t}, \tag{142}$$

which, under stationarity conditions ($|\lambda_j| \leq 1;\ j = 1, \ldots, n$), shows that $a_j(t)$ evolves as a damped spiral in the complex plane with a characteristic damping rate γ_j and frequency ω_j.

We now decompose the eigenvector (i.e., normal mode) $\boldsymbol{\phi}_j$ as the sum of a real and an imaginary term:

$$\boldsymbol{\phi}_j = \boldsymbol{\phi}_j^R + i\boldsymbol{\phi}_j^I. \tag{143}$$

Then, noting that for $\mathbf{B}$ real, the normal modes occur in complex conjugate pairs (if they are complex at all), and the general evolution of a damped normal mode (i.e., $\gamma_j \leq 1$) can be described in a two-dimensional subspace spanned by $\boldsymbol{\phi}_j^R$ and $\boldsymbol{\phi}_j^I$ (see e.g., von Storch et al. 1995). That is, the evolution of a damped mode occurs in a succession

$$\ldots \rightarrow \boldsymbol{\phi}_j^R \rightarrow -\boldsymbol{\phi}_j^I \rightarrow -\boldsymbol{\phi}_j^R \rightarrow \boldsymbol{\phi}_j^I \rightarrow \boldsymbol{\phi}_j^R \rightarrow \ldots \tag{144}$$

with a period of $2\pi/\omega_j$, each stage in (144) occurring a quarter of a cycle apart. Note that the time τ_j needed to reduce an initial *amplitude* $a_j(0)$ to $a_j(0)/exp(1)$ is referred to as the *e-folding time* and is given by:

$$\tau_j \equiv -\frac{1}{\ln(\gamma_j)}\,. \tag{145}$$

4.3 Estimation of POPs

The previous section emphasized the physical motivation behind the POPs analysis. If the system were *deterministic*, then the normal mode approach would be sufficient. However, the AR(1) representation is *stochastic* and, as such, must consider the effect of the error process.

In order to perform the POP analysis in practice, the system matrix $\mathbf{B}$ must be estimated. From (116) we see that a method-of-moments (MOM) estimator for $\mathbf{B}$ is

$$\hat{\mathbf{B}} = \hat{\mathbf{C}}_1^Z[\hat{\mathbf{C}}_0^Z]^{-1}\,, \tag{146}$$

where $\hat{\mathbf{C}}_0^Z$ is a MOM estimator, as shown in (65). Similarly, the (i,j)-th element of $\mathbf{C}_1^Z$ is given by,

$$c_1^Z(\mathbf{s}_i, \mathbf{s}_j) \equiv \mathrm{E}(Z(\mathbf{s}_i;t)Z(\mathbf{s}_j;t-1)), \tag{147}$$

so that the MOM estimator is

$$\hat{c}_1^Z(\mathbf{s}_i, \mathbf{s}_j) \equiv \frac{1}{T-1} \sum_{t=2}^{T} [Z(\mathbf{s}_i; t) - \hat{\mu}_z(\mathbf{s}_i; t)][Z(\mathbf{s}_j; t-1) - \hat{\mu}_z(\mathbf{s}_j; t-1)], \tag{148}$$

where possible choices for the mean estimator are discussed in Section 2.5.

The decomposition of $\hat{\mathbf{B}}$ then gives estimated eigenvectors $\hat{\boldsymbol{\phi}}_j$, adjoints $\hat{\boldsymbol{\psi}}_j$, and eigenvalues $\hat{\lambda}_j$, $j = 1, \ldots, n$. The estimated eigenvectors are sometimes referred to as *empirical* normal modes, analogous to the deterministic decomposition. It is then assumed, sometimes erroneously, that these empirical normal modes behave as we would expect the deterministic normal modes to behave. For example, for damped empirical modes (i.e., $\hat{\gamma}_j \leq 1$, where $\hat{\gamma}_j = |\hat{\lambda}_j|$) we expect the succession (144). However, in the presence of error, this relationship may not hold. To check if this relationship is valid in practice, a cross-spectral analysis is often performed between the estimates of the real component $a_j^R(t)$ and the imaginary component $a_j^I(t)$ of $a_j(t)$ which, according to the deterministic analysis, should vary coherently with a frequency ω_j and phase lag $\pi/2$, $a_j^R(t)$ lagging $a_j^I(t)$.

4.4 Application of POPs

As with all of the spatio-temporal methods used in the atmospheric sciences, POP analysis can be applied as either a diagnostic or a prognostic tool. Both of these applications will be briefly examined in the following subsections.

4.4.1 Diagnostic Applications of POPs

In a diagnostic mode, POPs are used to examine the oscillation properties and spatial structure of dynamical processes in the atmosphere. In this case, one looks at the estimated frequencies $\{\hat{\omega}_j\}$, amplitudes $\{\hat{\gamma}_j\}$, and e-folding times $\{\hat{\tau}_j\}$, as well as the amplitude time series $\{\hat{a}_j(t)\}$ and eigenvectors $\{\hat{\boldsymbol{\phi}}_j\}$. These quantities give insight into the physical meaning behind the empirical normal modes. Often, one first filters the data (usually in time) to focus on a particular atmospheric phenomenon. Essentially, this is an *ad hoc* method for removing the error, so that the deterministic interpretation is more tenable. Clearly, there is no guarantee

that such a noise-reduction scheme is optimal (almost certainly, it is not), but the issue has not been addressed in the literature. After filtering, it is hoped that the spatial patterns of the empirical normal modes can illustrate the spatial structure of the phenomenon of interest as it evolves in time. For example, von Storch et al. (1988) used POPs to consider the equatorial 30-60 day oscillation (Madden and Julian 1971).

4.4.2 Prognostic Applications of POPs

Since POPs have an inherent AR(1) dynamical structure, they are ideally suited for prognostic applications. Analogous to the deterministic case (135), it is easy to show from (110) that

$$a_j(t+1) = \lambda_j a_j(t) + \tilde{\eta}_j(t), \tag{149}$$

where

$$\tilde{\eta}_j(t) \equiv \frac{\boldsymbol{\psi}_j^* \boldsymbol{\eta}(t)}{d_j}. \tag{150}$$

The presence of the noise term clearly prevents the use of the deterministic normal mode results for prediction. However, this noise $\tilde{\eta}_j(t)$ is typically *ignored* in practice. The justification is that usually only one empirical normal mode is considered to be of physical importance, so after prefiltering in favor of this mode, the noise is assumed to be negligible. This is a tenuous assumption. In spite of these reservations, predictions can be obtained from

$$\hat{a}_j(t+1) = \hat{\gamma}_j e^{i\hat{\omega}_j} \hat{a}_j(t), \tag{151}$$

which then gives

$$\hat{\mathbf{Z}}_j(t+1) \equiv \hat{a}_j(t+1)\boldsymbol{\phi}_j\,, \tag{152}$$

and hence

$$\hat{\mathbf{Z}}(t+1) = \sum_{j=1}^{n} \hat{\mathbf{Z}}_j(t+1), \tag{153}$$

or some truncated version of (153). In the deterministic case, if we knew the location in the complex state space of the system at any given time, we could predict perfectly into the future. Clearly, the presence of noise limits the skill of any such approach. However, it is argued that

even in the presence of "unpredictable noise", such a scheme should be useful for short time leads (von Storch et al. 1995). It is not at all apparent that the noise $\{\tilde{\eta}_j(t)\}$ is necessarily "unpredictable". In fact, (150) shows that under Gaussian white noise assumptions for $\eta(t)$, $\tilde{\eta}_j(t)$ is simply a linear combination of Gaussian random variables, and so must itself be a Gaussian random variable, but *with* spatial dependence. We should then be able to use this dependence to increase our prognostic skill. This is the essence of time series and spatial prediction methodologies.

We note that the POP formulation and prediction methodology with a stationary model inherently assumes a decay in amplitude (since stationarity implies $\gamma_j \leq 1\,; j = 1, \ldots, n$). Thus, it is common to "respecify" the estimate of γ_j, by setting it equal to one. In this case, the amplitude does not change with time (i.e., a persistence forecast of amplitude is assumed). Then, the frequency ω_j takes on added importance since the choice of inital phase becomes critical. It is usually the case that $\mathbf{Z}(T)$ (i.e., the latest observations) are noisy, so that $a_j(T)$ is too noisy to use in the prediction. Thus, some form of noise reduction is applied. Typically, this entails projecting the data onto a limited set of the first K EOFs (to smooth the data in space) and to apply a time filter. The EOF projection is useful for other reasons as explained below. Examples of POP forecasting can be found in Xu and von Storch (1990) and von Storch and Xu (1990) as well as von Storch et al. (1995). Clearly, the noise reduction referred to above is *ad hoc*; optimal methods should increase the forecasting skill of POPs.

As mentioned above, the linear stationary nature of the POP methodology forces all oscillatory solutions to decay. In a diagnostic analysis this does not present a problem (and, actually, the decay rate can be useful information). However, it would seem to be quite inappropriate in the forecasting framework, particularly for the atmosphere, where most phenomena have amplitudes that are growing at some point in their evolution. Penland (1989) makes a strong case for using a large set of empirical normal modes in the forecast. In that case, constructive interference between the various empirical modes allows for the growth of certain multi-mode phenomena. Physically, this approach has much more appeal than the single mode approach.

While the POP approach has intuitive appeal in its deterministic form, perhaps the most glaring weakness of the approach is its treatment of error. In general, the error is simply ignored, leading to non-optimal predictions. In an attempt to deal with practical difficulties, a number of *ad hoc* "fix ups" have been used. We then ask the question, can POPs be reformulated in such a way as to account for the presence of error optimally? Kooperberg and O'Sullivan (1994) have recently addressed this issue with what they refer to as predictive oscillation patterns.

4.5 POPs on the EOF Basis

As mentioned in the previous section, before a POP analysis is conducted, the data can be projected onto a truncated set of EOFs in order to reduce the spatial dimension of the system. In that case, it is assumed that noise is then excluded from the analysis (although not optimally as we have shown in Section 2.4). The EOF expansion also improves the estimation from a practical numerical perspective, analogous to the use of principal components in linear regression (e.g., Draper and Smith 1981, p. 327-331). In the POP case, the estimate of $\mathbf{B}$ given in (146) contains the inverse of $\mathbf{C}_0^Z$, which could be very unstable if the dimension of n is large and the data are noisy. In that case, the smallest (and presumably, physically uninteresting) eigenvalues and associated eigenvectors of $\mathbf{C}_0^Z$ dominate the decomposition of $\mathbf{B}$ (Kooperberg and O'Sullivan 1994). Thus, by projecting the data onto the first K EOFs, we can reduce the spatial dimension from n to K and obtain a diagonal $\mathbf{C}_0^Z$ matrix that is quite easy to invert.

4.6 Extensions of POPs

In this section, some extensions of POPs are considered. Some, such as continuous time POPs, Complex POPs, and Cyclostationary POPs, have been considered in the literature. Others, such as POPs in the presence of measurement error, nonstationary POPs, and two-field POPs have not been considered, to the best of my knowledge.

4.6.1 POPs in Continuous Time

From a physical viewpoint, the time domain in the POP analysis should be considered as continuous. In that case, one must solve the appropriate Fokker-Planck equation to get a probabilistic solution to the stochastic differential equation describing the temporal evolution of the dynamical process of interest (e.g., Penland 1989). Penland (1989) and collaborators (Penland and Ghil 1993; Penland and Magorian 1993) have taken this approach and demonstrated its usefulness, particularly with regard to prognostic applications.

4.6.2 Complex POPs

Complex POP (CPOP) analysis was introduced by Bürger (1993) as an extension of conventional POP analysis. Bürger notes that while POP analysis is able to model *traveling oscillations*, they are unable to model *standing oscillations* (see Section 2.6.1 for an explanation of traveling and standing oscillations). In fact, he shows the inherent impossibility of modeling standing oscillations in first-order linear systems. So CPOP analysis is a natural extension to POP analysis, in many ways analogous to the relationship between EOFs and CEOFs. That is, CEOFs are able to detect traveling disturbances which cannot be detected by EOFs, and CPOPs can detect standing oscillations, which cannot be detected by POPs.

Just as for the CEOF analysis (Section 2.6.1), we write the system in terms of a new state process

$$\mathbf{W}(t+1) = \mathbf{G}\mathbf{W}(t) + \boldsymbol{\eta}(t), \tag{154}$$

where

$$\mathbf{W}(t) \equiv \mathbf{Z}(t) + i\tilde{\mathbf{Z}}(t), \tag{155}$$

and $\tilde{\mathbf{Z}}(t)$ is the Hilbert transform (Section 2.6.1) of $\mathbf{Z}(t)$. The CPOP analysis then proceeds in a similar manner to the POP analysis. However, since $\mathbf{G}$ is complex, its eigenvectors (i.e., the CPOPs) do not appear in complex conjugate pairs. von Storch et al. (1995) provides an excellent review of this approach.

4.6.3 Cyclostationary POPs

Traditional time series analysis techniques (including the vector autoregression case that is analogous to POPs) require an assumption of second-order stationarity (i.e., constant mean with autocorrelation depending only on time lag). This assumption clearly breaks down when the physical process under consideration has known cycles (i.e., solar influenced annual and semiannual cycles in atmospheric processes). In that case the mean and variance (at least) are also periodic. Traditionally, investigators remove these cycles, hoping that they then can satisfy the stationarity assumption (which typically, they cannot, at least with regard to the variance). However, from a statistical perspective, it makes sense to use the redundant information contained in the periodically correlated statistics optimally, rather than to remove it. An excellent discussion of the analysis of periodically correlated atmospheric time series can be found in Lund et al. (1995).

Blumenthal (1991) first published the idea of using periodically correlated statistics (i.e., cyclostationarity) in the POP framework. His approach is summarized in von Storch et al. (1995). In this case, the the cyclostationary process is written as

$$\mathbf{Z}(t,\tau+1) = \mathbf{A}(\tau)\mathbf{Z}(t,\tau) + \tilde{\boldsymbol{\eta}}(t,\tau), \tag{156}$$

where t and τ are integers such that t counts the cycles (e.g., years) and $\tau = 1,\ldots,m$ counts the position within each cycle (e.g., months). Then, $\mathbf{Z}(t,\tau+m) \equiv \mathbf{Z}(t+1,\tau)$ and $\mathbf{A}(\tau+m) \equiv \mathbf{A}(\tau)$. Thus we can write the cyclostationary POP model for $\tau = 1,\ldots,m$ as

$$\mathbf{Z}(t+1,\tau) = \mathbf{B}(\tau)\mathbf{Z}(t,\tau) + \boldsymbol{\eta}(t,\tau), \tag{157}$$

where

$$\mathbf{B}(\tau) \equiv \mathbf{A}(\tau+m-1)\mathbf{A}(\tau+m-2)\ldots\mathbf{A}(\tau), \tag{158}$$

and

$$\boldsymbol{\eta}(t,\tau) \equiv \sum_{j=0}^{m-1} \tilde{\boldsymbol{\eta}}(t,\tau+j)\,. \tag{159}$$

We then obtain a set of POPs (i.e., eigenvectors) for each τ.

4.6.4 POPs in the Presence of Measurement Error

The issue of measurement error in the model leading to POPs has not been addressed in the literature. If we assume that the true dynamical system of interest is given by the signal $\mathbf{Y}(t)$ in the presence of additive measurement error, then

$$\mathbf{Z}(t) = \mathbf{Y}(t) + \epsilon(t). \tag{160}$$

Thus, we can write the POP model in a state-space formulation

$$\mathbf{Z}(t) = \mathbf{Y}(t) + \epsilon(t) \tag{161}$$

$$\mathbf{Y}(t+1) = \boldsymbol{\beta}\mathbf{Y}(t) + \boldsymbol{\eta}(t). \tag{162}$$

This formulation has not been considered from a POP perspective but, because measurement error is so pervasive, it is an important area to pursue. It naturally would lead to a Kalman-filter implementation. Perhaps, if interested in a particular phenomenon, one could include one's prior belief about the process by specifying $\boldsymbol{\beta}$ in a Bayesian setting. This would lead to a hierarchical Bayesian analysis of the system (161),(162).

4.6.5 Two-Field POPs

We might also assume that there is a linear relationship between the process $\mathbf{Z}(t)$ and another process (of possibly different dimension) at a previous time $\mathbf{X}(t-1)$ given by:

$$\mathbf{Z}(t+1) = \boldsymbol{\beta}\mathbf{X}(t) + \boldsymbol{\nu}(t)\,, \tag{163}$$

where $\boldsymbol{\beta}$ is the two-field system matrix, and $\boldsymbol{\nu}(t)$ is some error process, independent of $\mathbf{X}(t)$. This model is essentially a two-field extension of POPs. We would then be interested in the decomposition of the matrix $\boldsymbol{\beta}$, which can be written as

$$\boldsymbol{\beta} = \mathbf{C}_1^{Z,X}[\mathbf{C}_0^X]^{-1}, \tag{164}$$

where

$$\mathbf{C}_1^{Z,X} \equiv \mathrm{E}[\mathbf{Z}(t+1)\mathbf{X}(t)'] \tag{165}$$

$$\mathbf{C}_0^X \equiv \mathrm{E}[\mathbf{X}(t)\mathbf{X}(t)']\,. \tag{166}$$

Note that this model does not include any dynamic structure on $\mathbf{X}(\cdot)$.

Neither the physical implications nor the statistical aspects of this model have been investigated. In particular, a comparison between this approach and the two-field spatio-temporal canonical correlation analysis (see below) would be interesting to pursue.

4.6.6 Non-Stationary POPs

As discussed previously, the stationary (in time) assumption in POP analysis forces damped oscillatory solutions, which present a problem for realistic prognostic applications. This suggests that one should investigate non-stationary vector AR models and their implications concerning growing modes (i.e., modes with increasing amplitudes) in POP-type analyses.

5 Space-Time Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) is a long-standing multivariate statistical technique (Hotelling 1936) that finds linear combinations of two sets of random variables, whose correlations are maximal. In the atmospheric sciences, CCA has been used in diagnostic climatological studies (e.g., Glahn 1968; Nicholls 1987; Barnett and Preisendorfer 1987), in the forecast of El Niño (Graham and Michaelsen 1987; Barnston and Ropelewski 1992), and the forecast of long-range temperature and precipitation (Barnston 1994). Bretherton et al. (1992) performed an intercomparison of methods for finding coupled patterns in climate data (including CCA and multivariate EOFs) and were supportive of a variant of CCA known (unfortunately) as Singular Value Decomposition (SVD). This is related to, but not the same as, the well-known matrix algebra technique of the same name. Some authors (e.g., Cherry 1994) have pointed out that there are difficulties with the interpretation of CCA and SVD results, particularly the tendency for the methods to produce spurious spatial patterns.

In the following, we shall review the traditional two-field space-time CCA approach used in the atmospheric sciences, briefly describe its estimation, examine the Singular Value Decomposition relative of CCA, consider the one-field time-lagged CCA approach, which is related to the

minimum/maximum autocorrelation factor (MAF) technique in statistics (Shapiro and Switzer 1989) and, finally, consider two extensions of CCA that have not been examined previously, CCA in the presence of measurement error and CCA in continuous space.

5.1 Two-Field Spatial-Temporal CCA

We assume that in addition to the process $Z(\mathbf{s};t)$ we are given another related process $X(\mathbf{s};t)$ with a possibly different spatial domain, but the same temporal domain. We further assume discrete space and time and zero means:

$$\mathrm{E}[\mathbf{Z}(t)] = \mathbf{0}_{n\times 1} \tag{167}$$

$$\mathrm{E}[\mathbf{X}(t)] = \mathbf{0}_{m\times 1}, \tag{168}$$

where $\mathbf{X}(t) \equiv (X(\mathbf{r}_1;t),\ldots,X(\mathbf{r};t))'$ and $\mathbf{Z}(t)$ is defined as before in (11). Furthermore, we define the covariances

$$\mathbf{C}_0^Z \equiv \mathbf{E}[\mathbf{Z}(t)\mathbf{Z}(t)']_{n\times n} \tag{169}$$

$$\mathbf{C}_0^X \equiv \mathbf{E}[\mathbf{X}(t)\mathbf{X}(t)']_{m\times m} \tag{170}$$

$$\mathbf{C}_0^{Z,X} \equiv \mathbf{E}[\mathbf{Z}(t)\mathbf{X}(t)']_{n\times m}, \tag{171}$$

which are invariant in time. We then define linear combinations of each data field

$$a_k(t) = \boldsymbol{\phi}_k'\mathbf{Z}(t) \tag{172}$$

$$b_k(t) = \boldsymbol{\psi}_k'\mathbf{X}(t), \tag{173}$$

where $k = 1,\ldots,\min\{m,n\}$. Now, define the k-th *canonical correlation* as

$$r_k \equiv \mathrm{corr}[\boldsymbol{\phi}_k'\mathbf{Z}(t), \boldsymbol{\psi}_k'\mathbf{X}(t)] \tag{174}$$

$$= \frac{\boldsymbol{\psi}_k'\mathbf{C}_0^{Z,X}\boldsymbol{\psi}_k}{(\boldsymbol{\phi}_k'\mathbf{C}_0^Z\boldsymbol{\phi}_k)^{1/2}(\boldsymbol{\psi}_k'\mathbf{C}_0^X\boldsymbol{\psi}_k)^{1/2}}. \tag{175}$$

The first pair of canonical variables are defined as the set of linear combinations $a_1(t)$ and $b_1(t)$ for $\{t = 1,\ldots,T\}$ that maximize the correlation (175) and have unit variance. The second pair of canonical variables are then the linear combinations $a_2(t)$ and $b_2(t)$ that are uncorrelated

with $a_1(t)$ and $b_1(t)$, have unit variance, and maximize (175). Then, the k-th set of canonical variables are the linear combinations $a_k(t)$ and $b_k(t)$ that are uncorrelated with the previous $k-1$ canonical pairs, have unit variance, and maximize (175).

Initially, let $k = 1$ and note that since $\mathbf{C}_0^Z$ and $\mathbf{C}_0^X$ are positive definite, they can be written as

$$\mathbf{C}_0^Z = (\mathbf{C}_0^Z)^{1/2}(\mathbf{C}_0^Z)^{1/2} \tag{176}$$

$$\mathbf{C}_0^X = (\mathbf{C}_0^X)^{1/2}(\mathbf{C}_0^X)^{1/2}. \tag{177}$$

Then, we can write

$$r_k^2 = \frac{[\tilde{\boldsymbol{\phi}}_1'(\mathbf{C}_0^Z)^{-1/2}\mathbf{C}_0^{Z,X}(\mathbf{C}_0^X)^{-1/2}\tilde{\boldsymbol{\psi}}_1]^2}{(\tilde{\boldsymbol{\phi}}_1'\tilde{\boldsymbol{\phi}}_1)(\tilde{\boldsymbol{\psi}}_1'\tilde{\boldsymbol{\psi}}_1)}, \tag{178}$$

where

$$\tilde{\boldsymbol{\phi}}_1 \equiv (\mathbf{C}_0^Z)^{1/2}\boldsymbol{\phi}_1 \tag{179}$$

$$\tilde{\boldsymbol{\psi}}_1 \equiv (\mathbf{C}_0^X)^{1/2}\boldsymbol{\psi}_1 \tag{180}$$

are normalized weights. The problem is now reduced to finding $\tilde{\boldsymbol{\phi}}_1$ and $\tilde{\boldsymbol{\psi}}_1$ that maximize (178). Note that $(\mathbf{C}_0^Z)^{-1/2}\mathbf{C}_0^{Z,X}(\mathbf{C}_0^X)^{-1/2}$ is not symmetric, which means that a singular value decomposition is needed. It can be shown (e.g., Johnson and Wichern 1992, p. 463) that r_1^2 is the largest singular value of

$$(\mathbf{C}_0^Z)^{-1/2}\mathbf{C}_0^{Z,X}(\mathbf{C}_0^X)^{-1/2}, \tag{181}$$

and $\tilde{\boldsymbol{\phi}}_1$ and $\tilde{\boldsymbol{\psi}}_1$ are the left and right singular vectors corresponding to r_1^2, respectively (see Section 4.1). Then, we can write

$$\boldsymbol{\phi}_1 \equiv (\mathbf{C}_0^Z)^{-1/2}\tilde{\boldsymbol{\phi}}_1 \tag{182}$$

$$\boldsymbol{\psi}_1 \equiv (\mathbf{C}_0^X)^{-1/2}\tilde{\boldsymbol{\psi}}_1. \tag{183}$$

We can also obtain the time series of canonical variables (for $t = 1, \ldots, T$):

$$a_1(t) = \boldsymbol{\phi}_1'\mathbf{Z}(t) \tag{184}$$

$$b_1(t) = \boldsymbol{\psi}_1'\mathbf{X}(t). \tag{185}$$

In general, $\tilde{\boldsymbol{\phi}}_k$ and $\tilde{\boldsymbol{\phi}}_k$ are the left and right singular vectors, respectively, associated with the k-th singular value r_k^2 from the singular value decomposition of (181). Then, $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$ can be obtained analogous to (182) and (183), as well as $a_k(t)$ and $b_k(t)$ analogous to (184) and (185).

We can also examine the correlation between the canonical variable time series and the original data. We obtain

$$\begin{aligned} r[a_k(t), \mathbf{Z}(t)] &\equiv \text{corr}[a_k(t), \mathbf{Z}(t)] && (186) \\ &= \boldsymbol{\phi}_k' \mathbf{C}_0^Z [\text{diag}(\mathbf{C}_0^Z)]^{-1/2} && (187) \\ r[b_k(t), \mathbf{X}(t)] &\equiv \text{corr}[b_k(t), \mathbf{X}(t)] && (188) \\ &= \boldsymbol{\psi}_k' \mathbf{C}_0^X [\text{diag}(\mathbf{C}_0^X)]^{-1/2}, && (189) \end{aligned}$$

where $[\text{diag}(\mathbf{C}_0^Z)]^{-1/2}$ is an $n \times n$ diagonal matrix with (i,i)-th element given by $\{\text{var}[Z(\mathbf{s}_i;t)]\}^{-1/2}$, and $[\text{diag}(\mathbf{C}_0^X)]^{-1/2}$ is an $m \times m$ diagonal matrix with (i,i)-th element given by $\{\text{var}[X(\mathbf{r}_i;t)]\}^{-1/2}$.

Equation (187) and (189) are known as the k-th left and right *homogeneous correlation maps*, respectively. These are maps in the sense that the i-th value of the vector is identified with the i-th location in Euclidean space. Similarly, we can define the k-th left and right *heterogeneous correlation maps*, respectively, as

$$\begin{aligned} r[a_k(t), \mathbf{X}(t)] &\equiv \text{corr}[a_k(t), \mathbf{X}(t)] && (190) \\ &= \boldsymbol{\phi}_k' \mathbf{C}_0^{X,Z} [\text{diag}(\mathbf{C}_0^X)]^{-1/2} && (191) \\ r[b_k(t), \mathbf{Z}(t)] &\equiv \text{corr}[b_k(t), \mathbf{Z}(t)] && (192) \\ &= \boldsymbol{\psi}_k' \mathbf{C}_0^{Z,X} [\text{diag}(\mathbf{C}_0^Z)]^{-1/2}. && (193) \end{aligned}$$

These maps represent how well the observations in one field can be explained by the k-th canonical variable from the other field.

5.2 Estimation of CCA

For the singular value decomposition, we need estimates of $\mathbf{C}_0^Z$, $\mathbf{C}_0^X$, and $\mathbf{C}_0^{Z,X}$. As with EOFs and POPs, estimation is performed by the method of moments:

$$\hat{\mathbf{C}}_Z \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}(t)\mathbf{Z}(t)' \tag{194}$$

$$\hat{\mathbf{C}}_X \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}(t)\mathbf{X}(t)' \tag{195}$$

$$\hat{\mathbf{C}}_{Z,X} \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}(t)\mathbf{X}(t)', \tag{196}$$

where both the $Z(\cdot)$ and $X(\cdot)$ processes are assumed to have zero means. Possible estimates of the mean are described in Section 2.5. Then, the estimated singular values and singular vectors are obtained from the numerical singular value decomposition of

$$(\hat{\mathbf{C}}_0^Z)^{-1/2}\hat{\mathbf{C}}_0^{Z,X}(\hat{\mathbf{C}}_0^X)^{-1/2}. \tag{197}$$

The CCA from these estimated matrices is often unsatisfactory because the covariance matrix estimates are noisy when estimated with sample sizes common in the atmospheric sciences. To compensate for this, the data are often projected onto a truncated set of EOFs before applying the singular value decomposition. This is the same technique suggested for use prior to a POP analysis (see Section 4.4.2). As in that case, we get a benefit from the reduction of spatial dimension (clearly necessary if $\min\{n, m\} > T$, which would otherwise imply a singular covariance matrix) and computational stability. As is always the case with truncated EOFs, there is some question about the appropriate number of EOFs to retain, and a substantial literature exists to help make this determination (e.g., Preisendorfer 1988).

5.3 Modifications of CCA

This section describes several modifications to CCA. In particular, there is a brief description of the popular Singular Value Decomposition relative of CCA, a description of the time-lagged version of CCA and its relationship to MAFs in statistics, as well as a brief exam-

ination of the previously unexamined issues of CCA in the presence of measurement error and CCA in continuous space.

5.3.1 Singular Value Decomposition Relative of CCA

Bretherton et al. (1992) popularized a close relative of CCA which they unfortunately call Singular Value Decomposition, and which we will refer to as SVD/CCA. In essence, SVD/CCA consists of finding $\breve{\phi}_k$ and $\breve{\psi}_k$ such that if

$$\breve{a}_k(t) = \breve{\phi}_k' \mathbf{Z}(t) \tag{198}$$

$$\breve{b}_k(t) = \breve{\psi}_k' \mathbf{X}(t), \tag{199}$$

then

$$c_k \equiv \mathrm{cov}[\breve{a}_k(t), \breve{b}_k(t)] \tag{200}$$

$$= \breve{\phi}_k' \mathbf{C}_0^{Z,X} \breve{\psi}_k \tag{201}$$

is maximized given uncorrelatedness with the $(k-1)$ previous values as well as the additional constraints

$$\breve{\phi}_k' \breve{\phi}_k = \breve{\psi}_k' \breve{\psi}_k = 1\,; \quad k = 1, \ldots, n\,. \tag{202}$$

Clearly, it is these constraints that account for the difference between SVD/CCA and CCA. Cherry (1994) discusses some of the implications of these additional constraints.

Then, the problem can be restated as one of finding the $\breve{\phi}_k$ and $\breve{\psi}_k$ that are, respectively, the right and left singular vectors from the singular value decomposition of $\mathbf{C}_0^{Z,X}$, where c_k is the k-th singular value.

5.3.2 Time-Lagged CCA

When only one process (say $Z(\cdot)$) is considered, the CCA technique can be used to find the canonical correlation patterns between $\mathbf{Z}(t)$ and $\mathbf{Z}(t+\tau)$, for some time lag τ. Such an analysis could then be useful for prognostic applications. In this case, we would like the vectors ϕ_k and

ψ_k such that the marginal correlation between

$$a_k(t) = \phi_k' \mathbf{Z}(t), \tag{203}$$

and

$$b_k(t+\tau) = \psi_k' \mathbf{Z}(t+\tau) \tag{204}$$

is maximized. To do this we obtain the k-th left and right singular vectors from the singular value decomposition applied to

$$(\mathbf{C}_0^Z)^{-1/2} \mathbf{C}_\tau^Z (\mathbf{C}_0^Z)^{-1/2}, \tag{205}$$

where

$$\mathbf{C}_\tau^Z \equiv \mathrm{E}[\mathbf{Z}(t)\mathbf{Z}(t+\tau)'], \tag{206}$$

and where we have assumed $Z(\cdot)$ has zero mean. Analogous to (182) and (183), to obtain ϕ_k and ψ_k we have to weight the eigenvectors obtained from this singular value decomposition by premultiplying them with $(\mathbf{C}_0^Z)^{-1/2}$.

The time-lagged CCA approach outlined here is similar to the POP analysis (Section 4) and to the minimum/maximum autocorrelation factor (MAF) approach in statistics (Shapiro and Switzer 1989; Cressie and Helterbrand 1994). In the MAF case, one is interested in the eigenvectors proportional to those obtained from the singular value decomposition of

$$(\mathbf{C}_0^Z)^{-1} \mathbf{V}, \tag{207}$$

where the matrix $\mathbf{V}$ is the first-difference correlation matrix:

$$\begin{aligned} \mathbf{V} &\equiv \mathrm{E}[(\mathbf{Z}(t) - \mathbf{Z}(t+1))(\mathbf{Z}(t) - \mathbf{Z}(t+1))'] && (208) \\ &= 2\mathbf{C}_0^Z - \mathbf{C}_1^Z - \mathbf{C}_1^{Z\prime} && (209) \\ &= 2\mathbf{C}_0^Z \left[\mathbf{I} - (\mathbf{C}_0^Z)^{-1} \left(\frac{\mathbf{C}_1^Z + (\mathbf{C}_1^Z)'}{2} \right) \right], && (210) \end{aligned}$$

where we have assumed temporal invariance in obtaining (210). Proportionality factors are used to ensure that the MAFs (i.e., $\alpha_k(t)$, where $\alpha_k(t) \equiv \phi_k' \mathbf{Z}(t)$; $k = 1, \ldots, n$) have unit variance and positive correlation with time.

When considering POPs, time-lagged CCA (for $\tau = 1$), and MAFs, we note that we must perform the singular value decomposition on the following matrices, respectively:

$$\mathbf{C}_1^Z(\mathbf{C}_0^Z)^{-1} \tag{211}$$

$$(\mathbf{C}_0^Z)^{-1/2}\mathbf{C}_1^Z(\mathbf{C}_0^Z)^{-1/2} \tag{212}$$

$$(\mathbf{C}_0^Z)^{-1}\left[\frac{\mathbf{C}_1^Z + (\mathbf{C}_1^Z)'}{2}\right]. \tag{213}$$

Clearly, these three matrices are similar in that they include some form of the lag-one covariance matrix and the inverse of the lag-zero covariance matrix. Furthermore, all three are square matrices but are nonsymmetric and thus, require the singular value decomposition. It is straightforward to show (using the respective eigen-equations) that the singular values are the same for the decomposition of (211) and (212) and that the singular vectors from the CCA decomposition (212) are equivalent to those from the POP decomposition (211) scaled by the matrix $(\mathbf{C}_0^Z)^{-1/2}$. The relationship between the eigenvalues and eigenvectors of the MAF matrix (213) and the POP and CCA matrices is more complicated. However, the MAF approach does have one clear advantage. Note that although $\mathbf{C}_1^Z$ is nonsymmetric, the matrix $\mathbf{C}_1^Z + (\mathbf{C}_1^Z)'$ is symmetric. Then, since the MAF approach is invariant to affine transformations (see Shapiro and Switzer 1989), a symmetric matrix can be obtained in (213) by first projecting the data onto its EOF basis (so that the lag-zero covariance matrix is the identity matrix). In this case, the symmetry implies that the singular vectors associated with $\mathbf{Z}(t)$ and $\mathbf{Z}(t+\tau)$ are the same (i.e., $\psi_k = \phi_k; k = 1, \ldots, n$), so we are able to consider the linear combinations $\phi_k'\mathbf{Z}(t)$ and $\phi_k'\mathbf{Z}(t+\tau)$, which is a useful feature. Additional investigation of the similarities and differences of these approaches would be beneficial.

5.4 CCA in the Presence of Measurement Error

Although the effect of sampling variability in CCA is well-known in the statistics literature (e.g., Johnson and Wichern 1993, Section 10.4), to the best of my knowledge, the issue of measurement error in CCA has not been considered. Assume that $\mathbf{Z}(t)$ and $\mathbf{X}(t)$ can be

written as,

$$\mathbf{Z}(t) = \mathbf{Y}(t) + \boldsymbol{\epsilon}(t) \tag{214}$$

$$\mathbf{X}(t) = \mathbf{W}(t) + \boldsymbol{\gamma}(t), \tag{215}$$

where $\mathbf{Y}(t)$ and $\mathbf{W}(t)$ are the true physical processes, and $\boldsymbol{\epsilon}$ and $\boldsymbol{\gamma}$ are additive measurement error. Then the following covariance function relationships are obtained:

$$\mathbf{C}_0^Z = \mathbf{C}_0^Y + \mathbf{C}_0^\epsilon \tag{216}$$

$$\mathbf{C}_0^X = \mathbf{C}_0^W + \mathbf{C}_0^\gamma, \tag{217}$$

where we have used an assumption that the errors are uncorrelated with their respective physical processes. Furthermore, if we assume that the two error processes are uncorrelated with each other, then we obtain the relationship,

$$\mathbf{C}_0^{Y,W} = \mathbf{C}_0^{Z,X}. \tag{218}$$

Now, let

$$\check{a}_k(t) \equiv \check{\boldsymbol{\phi}}_k' \mathbf{Y}(t) \tag{219}$$

$$\check{b}_k(t) \equiv \check{\boldsymbol{\psi}}_k' \mathbf{W}(t). \tag{220}$$

Following the same methodology as before, it is straightforward to show that the maximization of the squared correlation between (219) and (220) is accomplished through the singular value decomposition of

$$(\mathbf{C}_0^Z - \mathbf{C}_0^\epsilon)^{-1/2} \mathbf{C}_0^{Z,X} (\mathbf{C}_0^X - \mathbf{C}_0^\gamma)^{-1/2}. \tag{221}$$

Then we may ask two questions:

- How does the singular value decomposition of (221) compare to the singular value decomposition of (181) (i.e., without measurement error)?
- How do we estimate the measurement error covariances in (221)?

The first question may be answered approximately by expanding $(\mathbf{C}_0^Z - \mathbf{C}_0^\epsilon)^{-1/2}$ and $(\mathbf{C}_0^X - \mathbf{C}_0^\gamma)^{-1/2}$ in a power series, and truncating. The second question can be answered if separate data are available for assessment of the measurement error. Alternatively, one could fit a parametric model to the covariances, as is commonly done in geostatistical studies (e.g., Cressie 1993, Section 2.4).

5.5 CCA in Continuous Space

To the best of my knowledge, no one has considered CCA in a continuous space framework, analogous to the Karhunen-Loève (K-L) EOF approach. Of course, from a physical perspective, most of the spatial processes we deal with in atmospheric science are continuous, with observations at certain locations within a given domain. From this point of view, most of the data are missing and those that are available are collected together in a multivariate vector of observations to which multivariate methods are applied.

Consider the continuous spatial, discrete temporal processes:

$$Z(\mathbf{s};t):\ \mathbf{s}\in D_Z,\ t\in\{1,\ldots,T\} \tag{222}$$

$$X(\mathbf{r};t):\ \mathbf{r}\in D_X,\ t\in\{1,\ldots,T\}. \tag{223}$$

We then define

$$a_k(t) \equiv \int_{D_Z}\phi_k(\mathbf{s})Z(\mathbf{s};t)\,d\mathbf{s} \tag{224}$$

$$b_k(t) \equiv \int_{D_X}\psi_k(\mathbf{r})X(\mathbf{r};t)\,d\mathbf{r}. \tag{225}$$

We look for $\phi_k(\cdot)$ and $\psi_k(\cdot)$ such that $\mathrm{corr}[a_k(t),b_k(t)]$ is maximized, subject to the usual CCA orthogonality conditions. Let

$$r_k \equiv \mathrm{corr}[a_k(t),b_k(t)] \tag{226}$$

$$= \mathrm{corr}\left[\int_{D_Z}\phi_k(\mathbf{s})Z(\mathbf{s};t)\,d\mathbf{s},\int_{D_X}\psi_k(\mathbf{r})Z(\mathbf{r};t)\,d\mathbf{r}\right] \tag{227}$$

$$= \frac{\int_{D_X}\int_{D_Z}\phi_k(\mathbf{s})\psi_k(\mathbf{r})c_0^{Z,X}(\mathbf{s},\mathbf{r})\,d\mathbf{s}\,d\mathbf{r}}{\left[\int_{D_Z}\int_{D_Z}\phi_k(\mathbf{s})\phi_k(\mathbf{s})c_0^{Z}(\mathbf{s},\mathbf{s})\,d\mathbf{s}\,d\mathbf{s}\right]^{1/2}\left[\int_{D_X}\int_{D_X}\psi_k(\mathbf{r})\psi_k(\mathbf{r})c_0^{X}(\mathbf{r},\mathbf{r})\,d\mathbf{r}\,d\mathbf{r}\right]^{1/2}}, \tag{228}$$

where

$$c_0^Z(\mathbf{s},\mathbf{s}) \equiv \mathrm{E}[Z(\mathbf{s};t)Z(\mathbf{s};t)] \tag{229}$$

$$c_0^X(\mathbf{r},\mathbf{r}) \equiv \mathrm{E}[X(\mathbf{s};t)X(\mathbf{s};t)] \tag{230}$$

$$c_0^{Z,X}(\mathbf{s},\mathbf{r}) \equiv \mathrm{E}[Z(\mathbf{s};t)X(\mathbf{r};t)], \tag{231}$$

and we have assumed temporal invariance.

Intuitively, we would expect that there is a form of K-L singular value decomposition formulation that could be applied to (228) to aid in the maximization. Further effort is necessary to demonstrate this.

6 Conclusion

In Sections 2-5 we have given the fundamentals of the EOF, PIP, POP, and CCA techniques as applied in the atmospheric sciences. We have seen that there are many physical and statistical considerations that must be made when applying these techniques to data. In particular, we must decide if the problem should be considered discrete or continuous, with measurement error or without, prognostic or diagnostic, or involve spatial prediction or smoothing. We also must consider what is being optimized and whether there should be a dynamical (i.e., temporally dependent) component, what estimation strategy is appropriate, what are the effects of sampling variability, and whether Bayesian (e.g., Kalman filters) ideas should be considered. Some of these issues have been considered while others have only been touched upon. In fact, it is clear that most of these issues should be considered simultaneously. However, in practice this is rarely done.

Scientists are typically interested in their own scientific problems and not in what they may perceive as arcane statistical issues. They generally want simple, fast, yet powerful tools to help them achieve their scientific goals. Unfortunately, methods that are easy to use are usually very limited in their ability to deal with complicated problems. The spatio-temporal methods outlined in this review provide some very powerful diagnostic and prognostic tools for dealing

with high-dimensional spatio-temporal data sets. Although these methods are generally easy to implement in their traditional forms, in many cases extra effort is required to ensure that they provide the optimal results for the problem at hand. It is hoped that this review has identified some of the areas where additional efforts should be focused.

References

Achatz, U., G. Schmitz, and K.-M. Greisiger, 1995: Principal interaction patterns in baroclinic wave life cycles. *J. Atmos. Sci.*, **52**, 3201- 3213.

Barnett, T.P., 1977: The principal time and space scales of the Pacific trade wind fields. *J. Atmos. Sci.*, **34**, 221-236.

Barnett, T.P., 1983: Interaction of the monsoon and Pacific trade wind system at interannual time scales. Part I: The equatorial zone. *Mon. Wea. Rev.*, **111**, 756-773.

Barnett, T.P., and R.W. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115**, 1825-1850.

Barnston, A.G., 1994: Linear statistical short-term climate predictive skill in the Northern hemisphere. *J. Climate*, **7**, 1513-1564.

Barnston, A.G., and C.F. Ropelewski,1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, **5**, 1316-1345.

Blumenthal, B., 1991: Predictability of a coupled ocean-atmosphere model. *J. Climate,* **4**, 766-784.

Braud, I., and Ch. Obled, 1991: On the use of empirical orthogonal function (EOF) analysis in the simulation of random fields. *Stochastic Hydrol. Hydraul.*, **5**, 125-134.

Bretherton, C.S., C. Smith, and J.M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, **5**, 541-560.

Buell, C.E., 1972: Integral equation representation for factor analysis. *J. Atmos. Sci.*, **28**, 1502-1505.

Buell, C.E., 1975: The topography of empirical orthogonal functions. *Preprints Fourth Conf. Prob. Stats. Atmos. Sci.*, American Meteorological Society, 188-193.

Bürger, G., 1993: Complex principal oscillation pattern analysis. *J. Climate*, **6**, 1972-1986.

Chatfield, C., 1989: *The Analysis of Time Series: An Introduction*, Fourth Edition, Chapman and Hall, 241pp.

Cherry, S., 1994: Some comments on singular value decomposition and canonical correlation analysis. *J. Climate*, (submitted for review).

Christenson, W.I., Jr., and R.A. Bryson, 1966: An investigation of the potential of component analysis for weather classification. *Mon. Wea. Rev.*, **94**, 697-709.

Cohen, A. and R.H. Jones, 1969: Regression on a random field. *J. of the Amer. Stat. Assoc.*, **64**, 1172-1182.

Cressie, N.A.C, 1993: *Statistics for Spatial Data, Revised Edition,* Wiley, 900pp.

Cressie, N. and J.D. Helterbrand, 1994: Multivariate spatial statistical models. *Geographical Systems*, **1**, 179-188.

Daley, R., 1995: Estimating the wind field from chemical constituent observations: Experiments with a one-dimensional extended Kalman filter. *Mon. Wea. Rev.*, **123**, 181-198.

Davis, R.E., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific ocean. *J. Physical Oceanography,* **6**, 249-266.

Draper, N.R., and H. Smith, 1981: *Applied Regression Analysis*, Second Edition, Wiley, 709pp.

Freiberger, W. and U. Grenander, 1965: On the formulation of statistical meteorology. *Review of the International Statistical Institute,* **33**, 59-86.

Glahn, H.R., 1968: Canonical correlation and its relationship to discriminant analysis and multiple regression. *J. Atmos. Sci.*, **25**,23-31.

Graham, N.E., and J. Michaelsen, 1987: An investigation of the El Nino-Southern Oscillation cycle with statistical models. 1. Predictor field characteristics. *J. Geophys. Res.*, **92**, 14251-14270.

Gray, M., 1981: On the stability of temperature eigenvector patterns. *J. Climate*, **1**, 273-281.

Grewal, M.S., and A.P. Andrews, 1993: *Kalman Filtering: Theory and Practice*, Prentice-Hall, 381pp.

Hasselmann, K., 1988: PIPs and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns. *J. Geophys. Res.*, **93**, 11015 - 11021.

Horel, J.D., 1984: Complex principal component analysis: theory and examples. *J. Clim. Appl. Meteor.*, **23**, 1660-1673.

Hotelling, H., 1933: Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417-441,498-520.

Hotelling, H., 1936: Relations between two sets of variables. *Biometrika,* **28**, 321-377.

Johnson, R.A., and D.W. Wichern, 1992: *Applied Multivariate Statistical Analysis*, Prentice-Hall, 641 pp.

Karl, T.R., A.J. Koscielny, and H.F. Diaz, 1982: Potential errors in the application of principal component (eigenvector) analysis to geophysical data. *J. Appl. Meteor.*, **21**, 1183-1186.

Kooperberg, C. and F. O'Sullivan, 1994: Predictive Oscillation Patterns: a synthesis of methods for spatial-temporal decomposition of random fields. *Technical Report No. 277, Department of Statistics, University of Washington, Seattle..*

Kutzbach, J.E., 1967: Empirical eigenvectors of sea level pressure, surface temperature and precipitation complexes over North America. *J. Appl. Meteor.*, **6**, 791-802.

Lawley, D.N., 1956: Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, **43**, 128-136.

Loève, M., 1963: *Probability Theory,* Van Nostrand Company, 685 pp.

Lorenz, E.N., 1956: Empirical orthogonal functions and statistical weather prediction. *Sci. Rept. No. 1, Statistical Forecasting Project*, MIT, 49 pp.

Lund, R.B., H.L. Hurd, P. Bloomfield, and R. Smith, 1995: Climatological time series with periodic correlation. *J. Climate*, **8**, 2787-2809.

Madden, R.A., and P.R. Julian, 1971: Detection of a 40-50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.*, **28**, 702-708.

Miller, R.N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.*, **51**, 1037-1056.

Nicholls, N. 1987: The use of canonical correlation to study teleconnections. *Mon. Wea. Rev.*, **115**, 393-399.

North, G.R., T.L. Bell, R.F. Cahalan, and F.J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110**, 699-706.

Obled, Ch., and J.D. Creutin, 1986: Some developments in the use of empirical orthogonal functions for mapping meteorological fields. *J. Clim. Appl. Meteor.*, **25**, 1189-1204.

Papoulis, A., 1965: *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, Inc., pp. 457-461.

Peagle, J.N., and R. B. Haslam, 1982: Statistical prediction of 500mb height field using eigenvectors. *J. Appl. Meteor.*, **21**, 127-138.

Penland, C., 1989: Random forcing and forecasting using principal oscillation pattern analysis. *Mon. Wea. Rev.*, **117**, 2165-2185.

Penland, C., and M. Ghil, 1993: Forecasting Northern Hemisphere 700-mb geopotential height anomalies using empirical normal modes. *Mon. Wea. Rev.*, **121**, 2355-2372.

Penland, C., and T. Magorian, 1993: Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *J. Climate,* **6**, 1067-1076.

Preisendorfer, R.W., 1988: *Principal Component Analysis in Meteorology and Oceanography*, Elsevier, 425 pp.

Preisendorfer, R.W., and T.P. Barnett, 1983: Numerical model-reality intercomparison test using small-sample statistics. *J. Atmos. Sci.*, **40**, 1884-1896.

Rao, C.R., 1973: *Linear Statistical Inference and its Applications,* Wiley.

Rasmusson, E.M., P.A. Arkin, and W.Y. Chen, 1981: Biennial variation in surface temperature over the United States as revealed by singular decomposition. *Mon. Wea. Rev.*, **109**, 587-598.

Richman, M.B., 1981: Obliquely rotated principal components: An improved meteorological map typing technique? *J. Appl. Meteor.*, **20**, 1145-1159.

Richman, M.B., 1986: Review article: rotation of principal components. *J. Climatol.*, **6**, 293-335.

Shapiro, D.E., and P. Switzer, 1989: Extracting time trends from multiple monitoring sites. *Technical Report No. 132*, Department of Statistics, Stanford University, Stanford, CA.

Trenberth, K.E., and W.-T. K. Shin, 1984: Quasi-biennial fluctuations in sea level pressures over the Northern Hemisphere. *Mon. Wea. Rev.*, **112**, 761-777.

von Storch, H., G. Bürger, R. Schnur, and J.-S. von Storch, 1995: Principal oscillation patterns: a review. *J. Climate,* **8**, 377-400.

von Storch, H., and G. Hannoshock, 1985: Statistical aspects of estimated principal vectors (EOFs) based on small sample sizes. *J. Clim. Appl. Meteor.*, **24**, 716-724.

von Storch, H., T. Bruns, I. Fischer-Bruns, and K.H. Hasselmann, 1988: Principal Oscillation Pattern analysis of the 30 to 60 day oscillation in a GCM. *J. Geophys. Res.*, **93**, 11022-11036.

von Storch, H., and J.-S. Xu, 1990: Principal oscillation pattern analysis of the tropical 30- to 60-day oscillation. Part I Definition of an index and its prediction. *Climate Dynamics*, 4,175-190.

Wallace, J.M., 1972: Empirical orthogonal representation of time series in the frequency domain. Part II: Application to the study of tropical wave disturbances. *J. Appl. Meteor.*, **11**, 893-900.

Wallace, J.M., and R.E. Dickinson, 1972: Empirical orthogonal representation of time series in the frequency domain. Part I: Theoretical considerations. *J. Appl. Meteor.*, **11**, 887-892.

Weare, B.C., and J.S. Nasstrom, 1982: Examples of extended empirical orthogonal function analysis. *Mon. Wea. Rev.*, **110**, 481-485.

Xu, J.-S., 1992: On the relationship between the stratospheric QBO and the tropospheric SO. *J. Atmos. Sci.*, **49**, 725-734.

Xu, J.-S., and H. von Storch, 1990: Predicting the state of the Southern Oscillation using principal oscillation pattern analysis. *J. Climate*, **3**, 1316-1329.

ON THE SEMIANNUAL VARIATION IN THE NORTHERN HEMISPHERE EXTRATROPICAL HEIGHT FIELD

A paper to appear in the *Journal of Climate*

Christopher K. Wikle and Tsing-Chang Chen

Abstract

A qualitative examination of the spatial distribution of the maximum semiannual oscillation (SAO) amplitudes in the Northern Hemisphere (NH) extratropical 500-hPa height field suggests that the SAO has a very dominant zonally asymmetric east-west structure. A comparison between the NH stationary eddies and the NH SAO shows that the NH midlatitude SAO can be explained almost entirely as being a result of the spatial and temporal asymmetries in the annual variation of the stationary eddies. It is also shown that the polar area SAO maxima over Siberia and Alaska are related to the east-west stationary eddy distribution. It is suggested that, ultimately, the mechanism for the SAO in the NH extratropics is simply a result of land-sea contrasts, similar to the mechanism proposed by van Loon (1967) for the Southern Hemisphere (SH) extratropical SAO. The chief difference, however, is that the NH extratropics are dominated by *east-west* land-sea contrasts due to the large continental land masses in the NH, while the SH land-sea contrast reflects the *north-south* differential heating between Antarctica and the surrounding ocean.

1 Introduction

The annual progression of the seasonal cycle in meteorological variables has been of interest throughout human history. Meteorologists began to document scientifically these seasonal cycles in the early part of the twentieth century [for a review of some of this early work see Hsu and Wallace (1976a,b)]. In particular, harmonic analysis techniques have been used to examine the first two harmonics of the seasonal cycle: the annual and semiannual cycle. In general, outside of the tropics and the Southern Hemisphere (SH) high latitudes, the semiannual cycle accounts for a much smaller percentage of seasonal variance than does the annual cycle. However, in many places in which the semiannual cycle is perceived to be weak, the amount of variance it explains in variables such as surface temperature can be critical, for example, to the growth stages of various crops. Unfortunately, a coherent discussion of the complete hemispheric spatial structure and the associated dynamical mechanisms of the Northern Hemisphere (NH) extratropical semiannual oscillation (SAO) has not appeared. This paper will show that the NH extratropical SAO exhibits an east-west structure and is generally governed by the spatio-temporal asymmetries in the seasonal variation of the NH stationary eddies.

The Southern Hemisphere extratropical SAO is well documented, and there exists a fairly complete explanation as to its dynamic mechanism. Schwerdtfeger (1960) first documented the existence of a semiannual harmonic in the high latitudes of the SH. Later, van Loon (1967) confirmed this oscillation and showed that it is essentially the result of the differences between the surface energy budget over land (i.e., Antarctica) and over the ocean. In other words, the SH SAO is largely due to the north-south land-sea contrast in the high southern latitudes. Meehl (1991) used more comprehensive datasets and GCM simulations to reexamine van Loon's proposed SH SAO mechanism. Although Meehl's GCM simulations were not quantitatively accurate, they were in qualitative agreement with the observations, and he used both the revised datasets and GCM simulations to verify the SH SAO mechanism as proposed by van Loon.

Unlike the SH SAO, there is no uniformly accepted explanation of the hemispheric-scale

dynamic mechanism for the NH SAO. There have been several studies that have illustrated possible mechanisms. In considering the SAO of surface pressure, Schwerdtfeger and Prohaska (1956) concluded that the primary cause of the semiannual pressure oscillation should be sought in the upper levels of the middle and subpolar latitudes. They hypothesized that such an oscillation could be based on the "different solar heating of different latitude belts," which in turn could lead to half-yearly variations in the zonal wind field. White and Wallace (1978), in looking at global temperature data, found that the semiannual cycle at lower latitudes (around 30°) fluctuates in an opposite manner over land and ocean. Lanzante (1983) found that most singularities in the North American extratropical 700-hPa height field could be explained by a semiannual east-west oscillation in the North Pacific. In extending his analysis to a hemispheric domain, Lanzante (1985) suggested that the cause of the semiannual oscillation in 700-hPa heights is related to climatological connections involving the Asiatic monsoon. Weickmann and Chervin (1988) looked at the observed and GCM-simulated annual and semiannual cycle in the global wind field. Their GCM simulations were only marginally successful in describing the observed SAO. However, based on the observed data, they hypothesized that the forcing for the SAO in the global wind field is due to zonally asymmetric variations of deep convection in equatorial regions. They also briefly mention that as the SH high latitude semiannual cycle is caused by the north-south zonally symmetric land-sea contrast, the NH high latitude semiannual cycle may be due to east-west land-sea contrasts. Although important, these studies do not provide a coherent hemispheric explanation of the mechanism responsible for the NH SAO.

The purpose of this note is to show that the NH extratropical SAO exhibits an asymmetric east-west structure and, more importantly, that this oscillation is largely governed by the asymmetric seasonal variation of NH stationary eddies. Thus, since the NH stationary eddies are primarily a result of east-west land-sea contrasts (i.e., differential heating and topography), the mechanism for the NH high latitude semiannual cycle is qualitatively similar to that for the SH extratropical semiannual cycle in which north-south land-sea contrasts are critical.

In order to examine the NH extratropical SAO, NH 5°-degree latitude/longitude gridded monthly geopotential heights were obtained from the National Center for Atmospheric Research (NCAR). We extracted data from 1961 to 1992. Since the data originated from the octagonal grid of the National Centers for Environmental Prediction (formerly the National Meteorological Center), there were no data south of 20°N. In addition, the analyses only considered data to 80°N. Although the study focuses on the 500-hPa level, data were extracted at the 300-hPa and 700-hPa level as well. In addition, gridded sea-level pressure data were extracted for the same time periods and spatial locations as the height data.

2 Diagnostic Analysis

2.1 The Average Extratropical Northern Hemisphere Semiannual Oscillation

In order to see the spatial variation of the NH SAO, the semiannual harmonic amplitude of the NH extratropical 500-hPa height field is shown in Fig. 1a. This figure includes five primary centers of semiannual cycle amplitude listed in order of decreasing amplitude: northern Siberia (70°N,110°E), western Alaska (65°N,160°W), western Pacific (35°N,165°E), the western United States (40°N,110°W), and the Strait of Gibraltar (35°N,5°W). In addition, the area north of about 60° – 65°N shows a strong SAO amplitude. Minor maxima occur over northeastern North America, the west Atlantic/Caribbean, and the Gulf of Oman. Perhaps the most interesting feature of this spatial variation is the east-west asymmetric structure in the midlatitudes and the more north-south structure in the high latitudes, with larger maxima in the hemisphere from 40°E eastward through 140°W.

The SAO phase distribution is shown in Fig. 1b. Note that the phase is depicted by the orientation of the line segments (north-south orientation indicates maxima on 1 January and 1 July; east-west orientation corresponds to maxima on 1 April and 1 October; the segment length is the same at all locations). It is clear that the areas with maximum SAO amplitudes generally have midwinter and midsummer peaks. The areas with smaller SAO amplitudes tend to have spring-fall peaks (e.g., the North Pacific transition area, the Pacific area west of Baja

California, the Gulf of Mexico, northeastern North America, and central Europe).

The pronounced zonally asymmetric distribution in the SAO amplitude field suggests that it would be useful to look at the stationary eddy pattern found by subtracting the zonal average height from the time mean height at each grid point. We will refer to this annual average "eddy" variable as $\bar{Z}_E$. Figure 2 shows the 500-hPa annual average stationary eddy pattern. There are two primary maxima (ridges), one located over western North America and the other extending from the central Atlantic through Europe to central Asia. Minima occur over northeastern North America and over the region extending from the eastern coast of Asia through the North Pacific. The locations of the SAO maxima shown in Fig. 1a are located away from the eddy centers.

2.2 Northern Hemisphere Midlatitude Semiannual Oscillation

To examine the relationship between the eddy locations and the midlatitude SAO we consider the deviation of the monthly average stationary eddies from the annual average eddy field. Specifically, we first take the monthly time mean height at each location, subtract the monthly zonal average height, and then subtract the annual average eddy component (Fig. 2) at each location. These monthly eddy deviations are denoted by Z_E. Now, consider the time series of Z_E and the semiannual harmonic at the center of maximum SAO amplitude in the western United States at 40°N,110°W (location "A"). Figure 3a shows a plot of the semiannual harmonic at this location in addition to a plot of the monthly Z_E time series at the same location. Clearly, there is a very strong agreement between these two time series. Similar plots (not shown) for the SAO maxima in the western Pacific and over the Strait of Gibraltar show roughly the same level of agreement between the SAO and the Z_E time series. Now, consider a location of relatively low SAO amplitude, such as north-central North America (50°N,95°W; location "B"). Figure 3b shows the SAO and Z_E time series at this location. A striking feature of this plot is the very clear annual oscillation in the eddies with only a very slight semiannual component. Similar plots (not shown) for other areas of low SAO amplitude also show that the eddy time series is dominated by a distinct annual cycle rather than a semiannual cycle.

This is even true at a location such as (50°N,140°E), which is positioned almost directly in the center of the negative anomaly residing off the east Asian coast (Fig. 2). It is then appropriate to consider how the stationary eddies might be inducing an SAO at some locations while not at others.

We hypothesize that the SAO in the NH midlatitudes is related to the asymmetric response of the atmospheric circulation to the annual variation of solar heating. This relationship is evident in the asymmetric spatial response of the stationary eddy pattern between winter and summer. Figure 4a shows the 500-hPa winter [December, January, February (DJF)] average deviations from the annual average eddy field (Fig. 2), while Fig. 4b shows the corresponding summer [June, July, August (JJA)] pattern. The spring (March, April, May) and fall (September, October, November) eddy deviation patterns (not shown) are much weaker and represent the transition between the winter and summer fields. As expected, since the NH jet stream is less intense in the summer, the summertime eddy deviations are generally weaker than those in the winter (i.e., asymmetry in time). Additionally, there are subtle but important differences in the locations of the centers of the stationary eddy deviations and the horizontal structures between the winter and summer (i.e., asymmetry in space). In particular, consider the eddy pair over North America. The positive anomaly off the western U.S. coast in the winter weakens in magnitude and shifts from a northwest-southeast orientation in the winter to a northeast-southwest orientation as it changes phase in the summer. Similarly, the negative anomaly over northeastern North America in the winter is elongated and shifts to the southwest as it reaches its maximum positive phase in the summer.

In order to see how the asymmetric response of the stationary eddies leads to a semiannual cycle in the midlatitudes, consider location A (40°N, 110°W) as identified previously. Figure 5a shows the longitude-time plot of the monthly average 500-hPa stationary eddy deviations (Z_E) at 40°N. The vertical line drawn in Fig. 5a at 110W corresponds to the eddy time series shown in Fig. 3a. Following this longitude from January through December, it is clear that the secondary positive peak in the summer is due to the westward extension of the eddy anomaly

located to the east of location A. If location A were placed several degrees westward of 110°W (say, at 135°W), there would not be a secondary positive peak in the stationary eddy field since the westward extension of the northeastern North American positive summer eddy is limited. Figure 5b shows the corresponding latitude-time plot at 110°W. Here, the horizontal line at 40°N corresponds to the Z_E time series at location A shown in Fig. 3a. Figure 5b shows clearly that the secondary positive eddy anomaly in the summer is the result of the northward extension of the positive anomaly centered at 35°N. The apparent westward and northward shift of the eddy shown in Fig. 5 is simply a result of the asymmetric response to seasonal heating in the northeastern North American summertime positive anomaly. Specifically, the eddy shift is due to the westward extension and northwest-southeast to northeast-southwest reorientation of this positive anomaly, along with a corresponding change in orientation of the summertime negative anomaly over the western United States. Similar plots (not shown) for the other midlatitude SAO amplitude maxima also show that they are a result of the extension of eddy anomalies from the south and east due to similar changes in orientation between winter and summer eddy pairs. It is then clear that the locations of maximum midlatitude SAO amplitude occur near the centers of the annual average eddies because a strong midlatitude SAO requires the intrusion of a nearby eddy of opposite sign.

2.3 Northern Hemisphere High Latitude Semiannual Oscillation

We now consider the semiannual cycle maxima shown in Fig. 1a that is located in the northern polar region. Specifically, consider the maximum located over northern Siberia (location "C"; 70°N,110°E) and the maximum located over western Alaska (location "D"; 65°N,160°W). Although the entire polar region north of 65°N shows a relatively strong semiannual cycle, the areas from 40°E eastward to 140°W have a substantially stronger SAO than the remaining polar area. To examine the relationship between the stationary eddies and the SAO at locations C and D, consider plots of the SAO and the monthly Z_E time series shown in Figs. 6a and 6b, respectively. The SAOs at both locations are roughly similar in amplitude and phase. There are, however, differences in the nature of the annual variation of the eddy deviation time series

at the two locations. Location C shows strong agreement between the SAO and Z_E series with the notable exception of a very weak maximum in the eddy series during the winter months. Similarly, location D shows good agreement between the two series except for a weak maximum in the summer portion of the eddy series. At both locations C and D, the semiannual cycle in Z_E is modulated by a strong annual cycle in Z_E. For instance, from Fig. 4 it is apparent that location C is under the influence of a negative anomaly in the winter and a positive anomaly in the summer. Alternatively, location D is under the influence of a positive anomaly in the winter and a negative anomaly in the summer. This 180° eddy phase shift between location C and D explains the large summer Z_E peak at location C and the large winter peak at location D.

Even though the annual cycle of Z_E dominates at locations C and D, there is still a noticeable semiannual cycle in the Z_E time series, which is in phase with the strong SAO. Figure 4 implies, and monthly Z_E spatial plots (not shown) reiterate, that the semiannual cycle in Z_E is largely due to the asymmetric spatio-temporal response of the small eddy in the Arctic Ocean north of Alaska, centered along the date line at 70°N in the summer (Fig. 4b). Although the eddy asymmetries contribute to the SAO at these locations, it is clear from the differences between the SAO and Z_E time series in Fig. 6 that these asymmetries in Z_E are not sufficient to explain the SAO in the subpolar region. Thus, it seems possible that the semiannual cycle of the zonal mean component may also contribute to the observed SAO in the polar region.

2.4 Vertical Structure

To examine the spatial distribution of the SAO in the vertical, we include the SAO amplitude and phase plots for the 300-hPa height field (Figs.7a,b), 700-hPa height field (Figs.7c,d), and sea-level pressure (SLP) field (Figs.7e,f) in addition to the 500-hPa SAO shown in Figs.1a,b. Although the SAO amplitude increases with height, the general structure of the amplitude and phase fields is similar throughout all levels. There are, however, notable differences. For instance, the 300-hPa SAO amplitude (Fig.7a) maximum over the Eastern Mediterranean Sea

is not present at the other levels. In addition, above 700-hPa the SAO phase fields differ over several limited spatial regions [e.g., western North Atlantic at 700-hPa (Fig.7d), and central Asia and the Sea of Japan at 300-hPa (Fig.7b)]. The most pronounced differences in SAO structure occur in the SLP field. In particular, the primary SLP SAO centers (e.g., the western Pacific, Alaska, the Strait of Gibraltar, northern Siberia) are shifted eastward relative to the upper levels. In addition, the Siberian and western United States SLP amplitude maxima (Fig. 7e) are relatively weaker than the other SLP maxima, and there is a strong SLP SAO amplitude center extending from the Hudson Bay across Greenland and into the central North Atlantic that is not present in the SAO amplitude at upper levels. There are also numerous small-scale differences in SAO phase between the SLP (Fig. 7f) and higher levels.

Although the general structural similarity in SAO phase and the locations of SAO amplitude centers in the vertical (particularly between the levels from 700-hPa to 300-hPa) suggests an equivalent barotropic structure in the NH SAO, the differences outlined above suggest that it may not be appropriate to extend such a characterization to the surface. Some of the differences in the SLP SAO are due to the effect of regional differences in near-surface processes. Others are surely due to data quality problems and SLP measurement deficiencies over complex and elevated terrain. In addition, we have demonstrated that the SAO is likely governed by the temporal and spatial asymmetries in the seasonal cycle of the stationary eddies. Thus, since Lau (1979) showed in longitude-height eddy cross sections that there is a relatively strong westward tilt from the surface up to 700-hPa at mid- and high latitudes, it is not unexpected that we should find the SAO centers in SLP shifted eastward relative to the SAO at upper levels.

We mention one additional feature of the SAO in SLP. The phase shown in Fig. 7f over western Alaska is approximately 90° out of phase from that reported by Hsu and Wallace (1976b, their Fig. 4). It is likely that this SLP SAO phase difference is due to the additional data (after the mid-1970s) used in our analysis. Specifically, it has been demonstrated that there was a significant climate regime shift in the North Pacific around 1976-77 (e.g., Trenberth

1990; Trenberth and Hurrell 1994; Graham 1994). This regime shift is particularly clear in the SLP data associated with the Aleutian low. In a recent study, van Loon et. al (1993) showed that a similar interdecadal variation in the SH atmospheric circulation led to an interdecadal variation in the SH high latitude SAO. Thus, it is likely that the interdecadal variability in the NH stationary eddies may force a similar interdecadal variability in the NH SAO.

3 Discussion

It is generally believed that the stationary eddies are the result of the atmosphere's response to topographic forcing and zonally asymmetric diabatic heating (e.g., Held 1983). Thus, it is clear that the atmospheric response to the annual variation in solar heating, particularly the seasonal land-sea temperature contrast related to the varying heat capacities of land and water, should not be symmetric. Indeed, it is well known that the winter and summer eddies have different characteristics (e.g., Wallace 1983). As shown above, these spatial and temporal asymmetries apparently induce a semiannual cycle in the NH midlatitudes. Thus, the east-west structure in the spatial distribution of the maximum NH SAO amplitude may be explained as the result of the atmospheric response to the east-west land-sea contrast. This then implies that the basic mechanism for the NH midlatitude SAO is the same as the mechanism for the SAO in the SH extratropics, as described by van Loon (1967). In the SH case, the north-south differential heating between the continent (Antarctica) and the surrounding ocean induces the SAO. The SAO in the NH midlatitudes is also due to land-sea contrasts, but the essential land-sea contrast (temperature and topography) has a general east-west structure due to the continental land masses. Consequently, the SH SAO shows a distinct north-south spatial structure, while the NH SAO shows a distinct east-west zonally asymmetric spatial structure.

As noted in section 2.3, the NH subpolar latitude SAO is not completely explained by the east-west asymmetries in the stationary eddy field. It is conjectured that the underlying subpolar SAO might be explained by the inclusion of the semiannual variation of the zonal mean component. In that case, the NH SAO may be at least partially induced by north-south

land-sea contrasts similar to (but much weaker than) the SH SAO. Unfortunately, analyzed height field data in the high latitudes are suspect, and therefore the immediate verification of this speculation is hampered. It is apparent, however, that the Siberia-Alaska region shows a much stronger SAO than other polar areas and that this stronger SAO is due to the modulating effects of the east-west asymmetries in the stationary eddy field. Thus, we can say that although the SAO in this region is governed by land-sea contrasts, it may contain contributions from both north-south and east-west differences.

4 Conclusions

Based on a qualitative examination of the spatial distribution of the maximum SAO amplitudes in the NH extratropical 500-hPa height field, we have concluded that this oscillation has a very dominant zonally asymmetric east-west structure. This in turn suggested that the stationary eddies might be a useful tool to explain this oscillation. A comparison between the stationary eddies and the SAO showed that the NH midlatitude SAO can be explained as almost entirely a result of the spatial and temporal asymmetries in the annual variation of the stationary eddies. It has also been shown that the polar area SAO maxima over Siberia and Alaska are related to the east-west stationary eddy distribution, but that such a relationship may not be sufficient to explain the oscillation. It has been suggested that, ultimately, the mechanism for the SAO in the NH extratropics is simply a result of land-sea contrasts, similar to the mechanism proposed by van Loon (1967) for the SH extratropical SAO. The chief difference, however, is that the NH extratropics are dominated by east-west land-sea contrasts due to the large continental land masses in the NH, while the SH land-sea contrast reflects the north-south differential heating between Antarctica and the surrounding ocean.

The suggested mechanism for the NH extratropical SAO was based on a diagnostic analysis of the geopotential height field. To test these hypotheses, one should make use of GCM experiments similar to Meehl's (1991) verification of the van Loon SAO mechanism in the SH. In turn, this problem would provide an important sensitivity test for the GCM. After

all, before we can accept GCM results related to interannual and interdecadal variability, we must be sure that the GCM can properly simulate the relatively simple components of the atmospheric seasonal cycle, namely the SAO. Experiments to date (e.g., Weickmann and Chervin 1988; Meehl 1991) have not demonstrated that the models are able to simulate the SAO with the desired level of accuracy. In addition, polar station data could be utilized to examine the possible underlying north-south structure in the polar region SAO, as well as the possible interdecadal variability of the NH SAO over western Alaska.

Acknowledgements

The research was sponsored by the U.S. Department of Energy, Office of Energy Research, Environmental Sciences Division, Office of Health and Environmental Research under the first author's appointment to the Graduate Fellowships for Global Change administered by Oak Ridge Institute for Science and Education. Additional support was provided by NSF Grant ATM-9416954. We wish to thank Dr. Richard Carlson for his discussion that led to this work, Ms. Susan Kiehne for her helpful comments on an early draft, and the anonymous reviewers who made several valuable suggestions that improved the manuscript.

References

Graham, N.E., 1994: Decadal-scale climate variability in the tropical and North Pacific during the 1970s and 1980s: Observations and model results. *Climate Dyn.*, **10**, 135-162.

Held, I.M., 1983: Stationary and quasi-stationary eddies in the extratropical troposphere: Theory. *Large-Scale Dynamical Processes in the Atmosphere,* B.J. Hoskins and R.P. Pearce, Eds., Academic Press, 127-168.

Hsu, C.-P. F., and J.M. Wallace, 1976a: The global distribution of the annual and semiannual cycles in precipitation. *Mon. Wea. Rev.,* **104,** 1093-1101.

Hsu, C.P. F., and J.M. Wallace, 1976b: The global distribution of the annual and semiannual cycles in sea level pressure. *Mon. Wea. Rev.,* **104,** 1597-1601.

Lanzante, J.R., 1983: Some singularities and irregularities in the seasonal progression of the 700 mb height field. *J. Climate Appl. Meteor.*, **22**, 967-981.

Lanzante, J.R., 1985: Further studies of singularities associated with the semiannual cycle of 700 mb heights. *Mon. Wea. Rev.*, **113**, 1372-1378.

Lau, N.-C., 1979: The observed structure of tropospheric stationary waves and the local balances of vorticity and heat. *J. Atmos. Sci.*, **36**, 996-1016.

Meehl, G.A., 1991: A reexamination of the mechanism of the semiannual oscillation in the southern hemisphere. *J. Climate*, **4**, 911-926.

Schwerdtfeger, W., 1960: The seasonal variation of the strength of the southern circumpolar vortex. *Mon. Wea. Rev.*, **88**, 203-208.

Schwerdtfeger, W., and F. Prohaska, 1956: The semiannual pressure oscillation, its cause and effects. *J. Meteor.*, **13**, 217-218.

Trenberth, K.E., 1990: Recent observed interdecadal climate changes in the Northern Hemisphere. *Bull. Amer. Meteor. Soc.*, **71**, 988-993.

Trenberth, K.E., and J.W. Hurrell, 1994: Decadal atmosphere-ocean variations in the Pacific. *Climate Dyn.*, **9**, 303-319.

Wallace, J.M., 1983: The climatological mean stationary waves: Observational evidence. *Large-Scale Dynamical Processes in the Atmosphere*, B.J. Hoskins and R.P. Pearce, Eds., Academic Press, 27-53.

Weickmann, K.M., and R.M. Chervin, 1988: The observed and simulated atmospheric seasonal cycle. Part I: Global wind field modes. *J. Climate*, **1**, 265-289.

White, G.H., and J.M. Wallace, 1978: The global distribution of the annual and semiannual cycles in surface temperature. *Mon. Wea. Rev.*, **106**, 901-906.

van Loon, H., 1967: The half-yearly oscillation in middle and high southern latitudes and the coreless winter. *J. Atmos. Sci.*, **24**, 472-486.

van Loon, H., J.W. Kidson, and A.B. Mullan, 1993: Decadal variation of the annual cycle in the Australian dataset. *J. Climate*, **6**, 1227-1231.

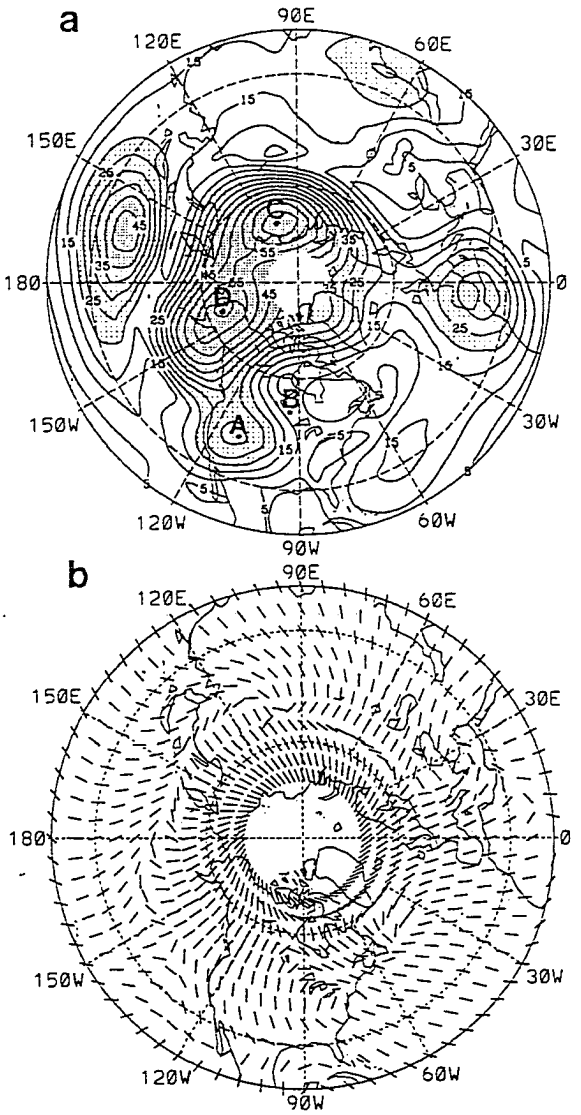

Figure 1 (a) Semiannual harmonic amplitude for the 500-hPa height field. The contour interval is 5 m. Values greater than 20 m are lightly shaded and values greater than 40 m are heavily shaded. (b) Semiannual harmonic phase line segments for the 500-hPa height field. Phase is indicated by the orientation of the line segments, with a segment oriented north-south corresponding to maxima on 1 January and 1 July, and an east-west orientation corresponding to maxima on 1 April and 1 October.

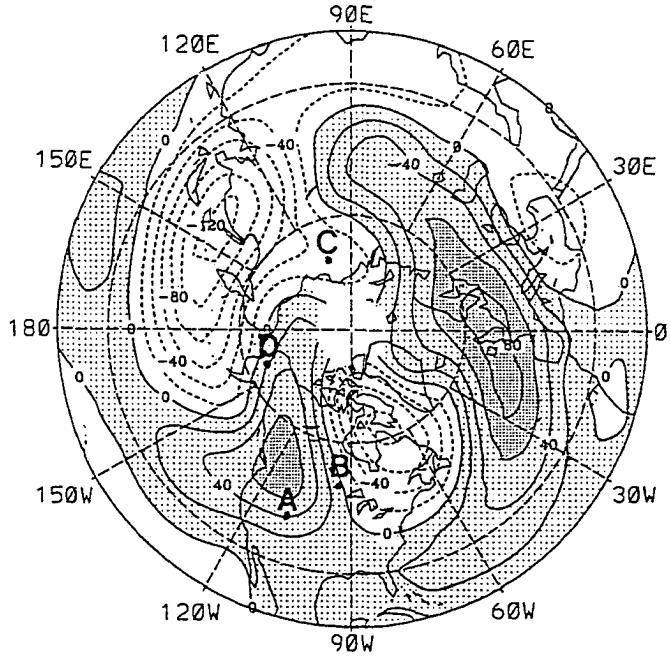

Figure 2 Annual average eddies for the 500-hPa height field, $\tilde{Z}_E$. The eddies are determined by subtracting the zonal mean from the time mean at each location. Contour interval is 20 m, with positive anomalies lightly shaded and anomalies greater than 60 m heavily shaded.

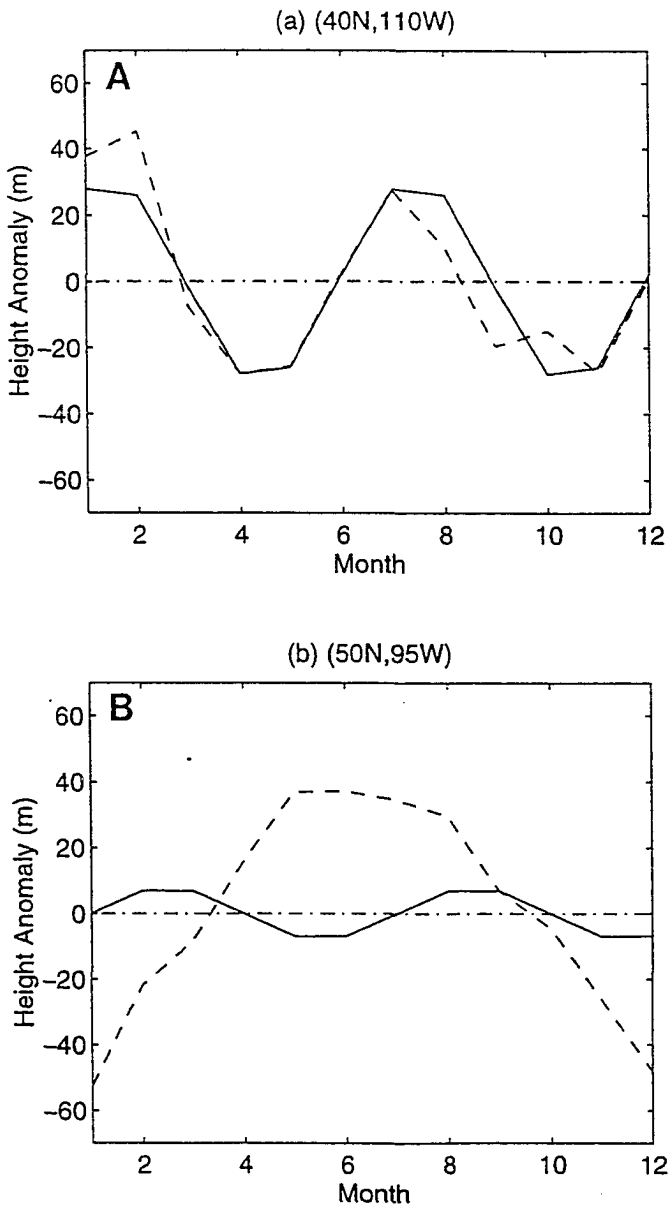

Figure 3 Monthly average time series (in meters) of the 500-hPa height field semiannual harmonic amplitude (solid line), and time series of 500-hPa monthly average deviations from the annual average stationary eddy field (dashed line) at (a) location A (40°N,110°W), and (b) location B (50°N,95°W).

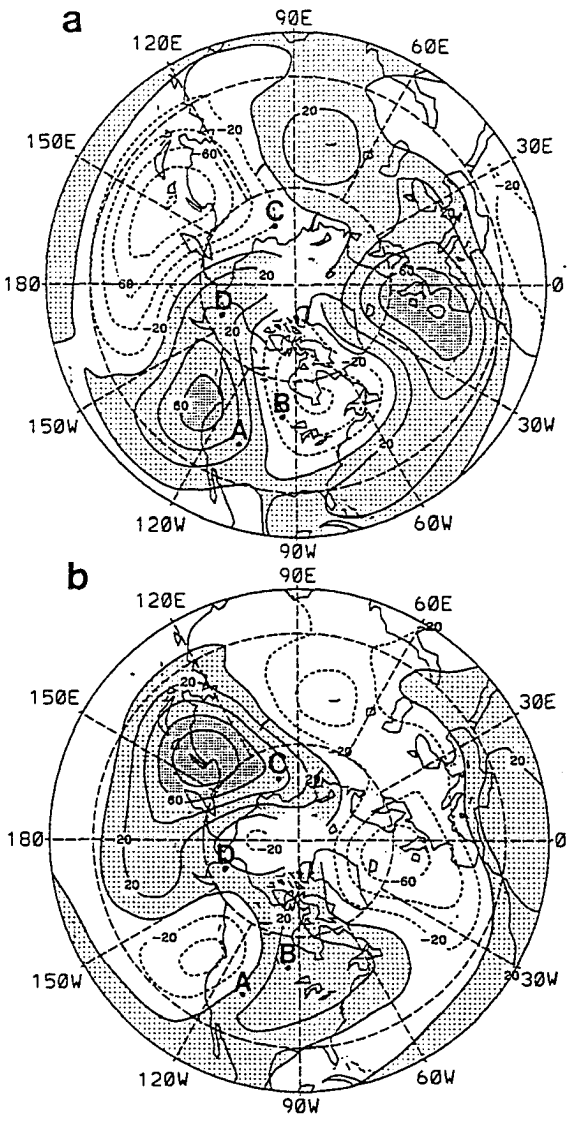

Figure 4 500-hPa geopotential height deviations from the annual average eddy field for (a) winter (DJF) and (b) summer (JJA). The contour interval is 20 m, and positive anomalies are shaded with heavier shading for positive anomalies greater than 60 m.

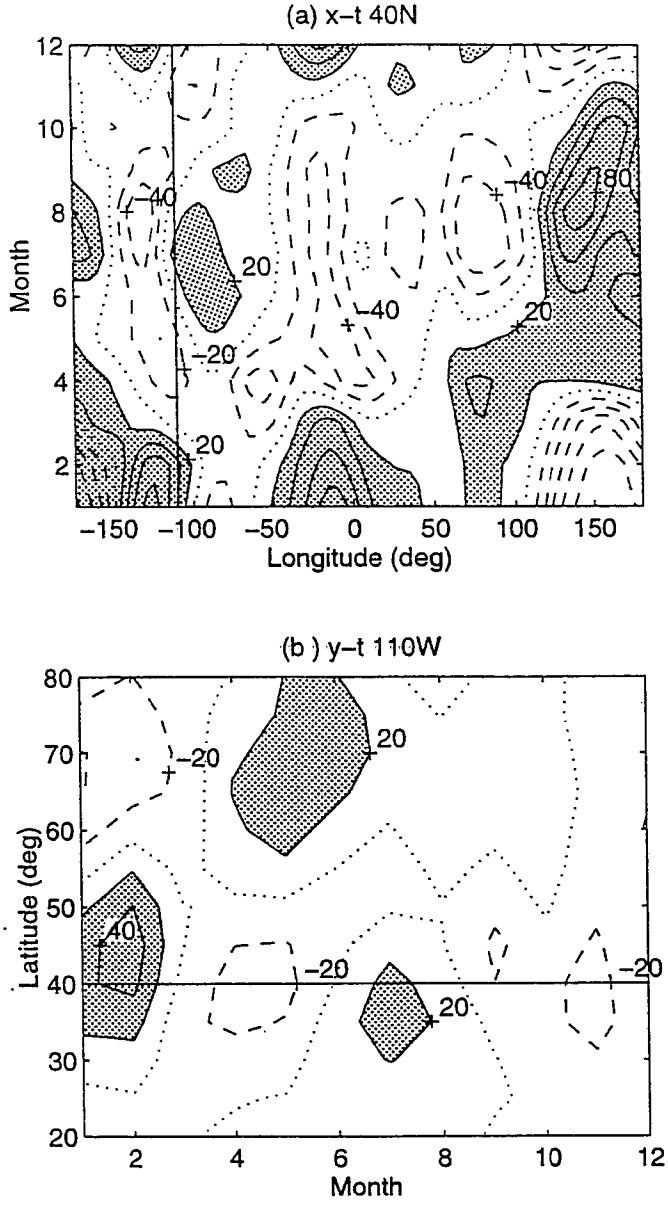

Figure 5 (a) Longitude-time (x-t) plot of Z_E at 40° N. (b) Latitude-time (y-t) plot of Z_E at 110°W. The vertical line at 110°W in (a) and the horizontal line at 40°N in (b) correspond to the Z_E time series in Fig. 3a. The contour interval is 20 m with solid lines for positive anomalies, dashed lines for negative anomalies, and a dotted line for the 0-meter line. Positive contours greater than 20 m are shaded.

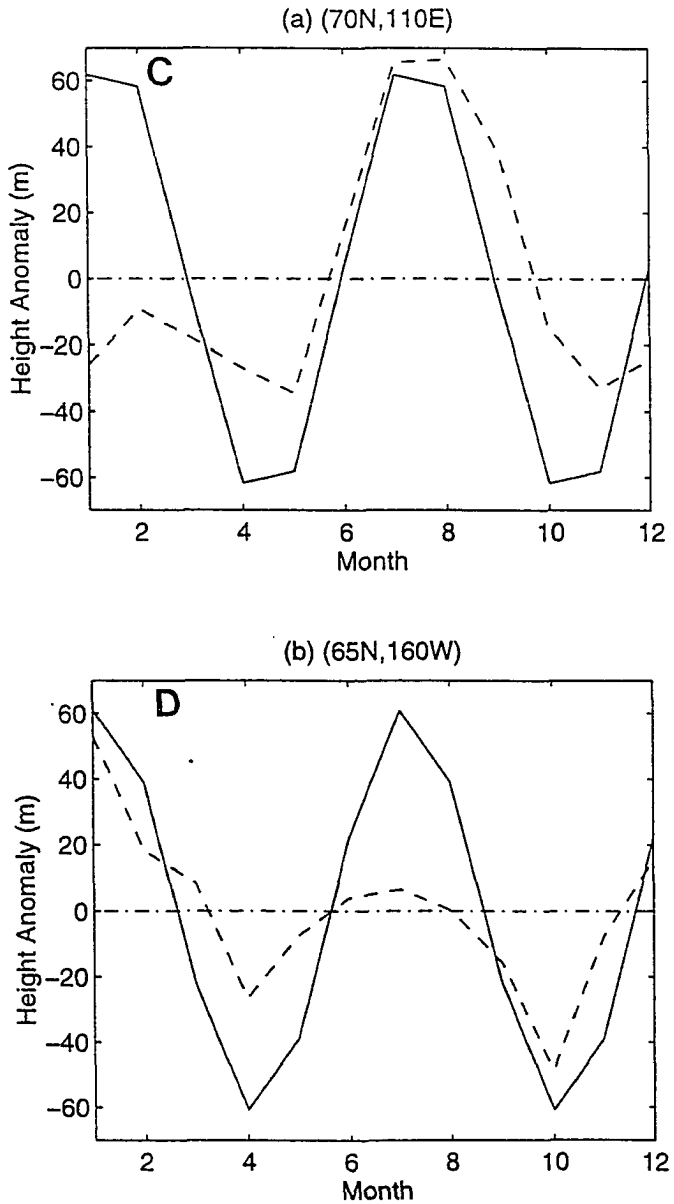

Figure 6 Monthly average time series (in meters) of the 500-hPa height field semiannual harmonic amplitude (solid line), and time series of 500-hPa monthly average deviations from the annual average eddy field (dashed line) at (a) location C (70°N,110°E), and (b) location D (65°N,160°W).

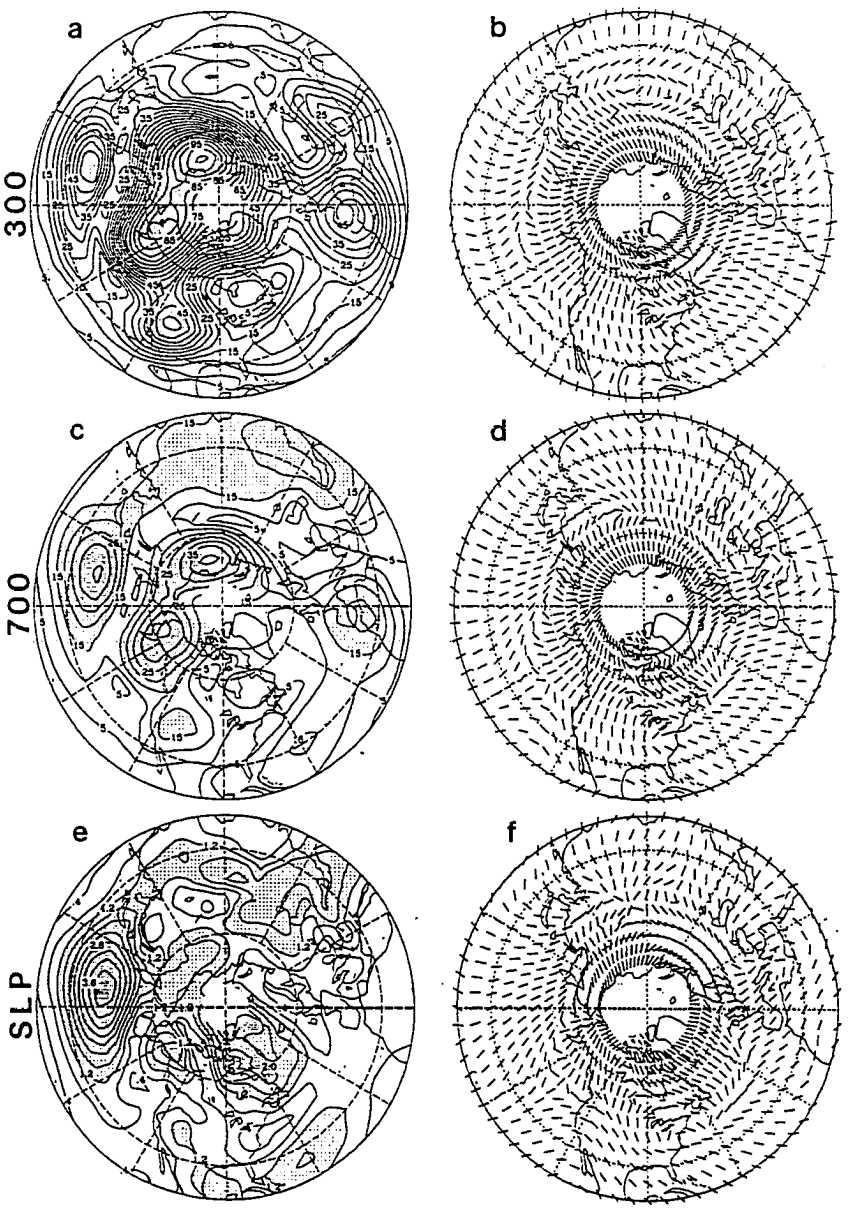

Figure 7 Semiannual harmonic amplitude for the (a) 300-hPa height field, (c) 700-hPa height field with a contour interval of 5 m, and (e) sea level pressure with a contour interval of 0.4-hPa. Semiannual harmonic phase for the (b) 300-hPa height field, (d) 700-hPa height field, and (f) sea level pressure with phase line segments as described for Fig. 1b.

SEASONAL VARIATION OF LOWER STRATOSPHERIC MIXED ROSSBY-GRAVITY WAVES OVER THE TROPICAL PACIFIC

A paper to be submitted to the

Journal of the Atmospheric Sciences

Christopher K. Wikle, Roland A. Madden, and Tsing-Chang Chen

Abstract

Lower stratospheric (70-hPa) wind data spanning 31 years from 1964 to 1994 were analyzed at four rawinsonde stations in the central/western Pacific. Traditional spectral and cross-spectral analysis led us to conclude that there is a significant signal with periods between 3 - 4.5 days, which we link with the dominant waves predicted by theory to have these periods, mixed Rossby-gravity waves (MRGWs). We then applied the seasonally varying spectral analysis method developed by Madden (1986) to study the average seasonal variation of these waves. The seasonally varying analysis suggested that there are significant twice-yearly maxima in MRGW activity, with peaks occurring in winter-early spring and in summer-early fall. Similar results were also shown by a cyclic spectral analysis. In addition, the seasonally varying mean-squared coherence between the u- and v-winds and the associated phase implied that there is convergence of horizontal momentum flux associated with these waves, and that the sign of that convergence is different during the two maxima. The cyclic spectral analysis also suggested that the frequency of the v-wind power is different during the two maxima. Assuming that these

spectral peaks are indicative of increased MRGW activity, then these frequency differences could be due to either the seasonal variation of the basic zonal state or to different MRGW forcing mechanisms (i.e., convective versus lateral excitation mechanisms).

1 Introduction

The foundation of the theory of equatorial waves was presented in the landmark paper of Matsuno (1966). Matsuno showed, using a simplified set of hydrodynamical equations, that equatorially confined wave solutions (among them the westward propagating mixed Rossby-gravity wave) were theoretically possible in tropical regions. Mixed Rossby-gravity waves (MRGWs) were first observed in rawinsonde data by Yanai and Maruyama (1966). As summarized by Andrews et al. (1987), subsequent studies have shown that MRGWs in the lower stratosphere are generally wave number 4 with westward phase propagation and periods in the range of 3-5 days. Furthermore, the MRGW phase tilt is westward with increasing height and the wave is expected to be observed when the mean zonal flow is westerly.

1.1 MRGW Forcing

Lindzen and Matsuno (1968) interpreted the westward tilting MRGWs as oscillations forced from below and thus suggested upward transport of momentum theoretically. Subsequently, Yanai and Hayashi (1969) used co-spectral analysis of the meridional wind and temperature and found a significant upward flux of wave energy at the tropopause level near the equator in the MRGW period range. They suggested that this implied a tropospheric origin of the stratospheric equatorial waves.

From the late 1960's to the present, there has not been a consensus as to the forcing mechanism of the MRGWs. Two schools of thought have generally dominated the literature. One theory concerns lateral forcing from the midlatitudes as originally proposed by Mak (1969). The other predominant school of thought considers latent heat of condensation from convection as a forcing mechanism. One convective forcing theory considers the interaction of cumulus

convection in a wave-CISK mechanism as proposed by Hayashi (1970). Another prominent convective forcing theory, proposed by Holton (1972), does not include a feedback mechanism from the excited motion to the heating. Itoh and Ghil (1988) combined the nonlinear wave-CISK and lateral forcing ideas to formulate a plausible theory of MRGW forcing. Recently, Goswami and Goswami (1991) and Emmanuel (1993) showed that so called wind-induced surface heat exchange (WISHE) models could produce realistic MRGWs.

Observational evidence exists which seems to validate both the lateral forcing (e.g., Zangvil and Yanai 1980; Yanai and Lu 1983; Magaña and Yanai 1995) and convective forcing (e.g., Nitta 1970; Zangvil and Yanai 1981; Hendon and Liebmann 1991) theories. Thus, since observational evidence exists to support both theories, it is likely that, as suggested by Itoh and Ghil (1988) and Dunkerton (1993), both mechanisms are important to the forcing of MRGWs. In fact, Hayashi and Golder's (1978) GCM simulation showed just that. They found that if midlatitude disturbances are eliminated, MRGWs appear in the stratosphere, probably due to the latent heat release from convection. However, the MRGWs were found to be significantly intensified by westward moving midlatitude disturbances which are found to propagate intermittently toward the equator (see also Magaña and Yanai 1995). These results support the thermal forcing theories and the lateral forcing theory. It was shown, however, that midlatitude lateral forcing was not sufficient to generate MRGWs when condensational heating was absent. Recently, it has been suggested (e.g., Dunkerton and Baldwin 1995) that, in addition to convective and lateral forcings, MRGWs can sometimes be excited by off equatorial tropical-depression disturbances.

1.2 Seasonal Variation of MRGWs

Although there have been numerous studies on the spectral characteristics of MRGWs and their potential forcing mechanisms, until recently little work has examined the seasonal variability of these waves. An early exception was Maruyama's (1969) finding that MRGWs appear predominantly when the absolute value of the zonal wind speed is decreasing in time. In addition, Hayashi and Golder (1980) considered the seasonal variation of MRGWs at the

tropopause level in the Geophysical Fluid Dynamics Laboratory (GFDL) general circulation model. They found that MRGWs attain a primary and secondary maxima around July and January, respectively. They suggest that although the semiannual variation may be related to the sun crossing the equator twice a year, the difference in amplitudes of the maxima suggest seasonal asymmetry in the zonal mean state, tropical cumulus convection, or the strength of midlatitude systems.

Hendon and Liebmann (1991) looked for the signature of 4-5 day period MRGWs in the tropical convection field across the Indian and Pacific oceans. By examining outgoing longwave radiation (OLR) data and gridded wind data, they found that antisymmetric fluctuations of tropical convection exhibit a spectral peak at 4-5 day periods only during the northern fall, within about 30 degrees longitude of the date line, with the peak occurring around 7.5° degrees latitude.

Dunkerton (1991b) showed that the descent of the quasi-biennial oscillation (QBO) easterlies is rapid during March-June and slow for July-February. Thus, since the descent of the QBO easterly wind regime is thought to be related to the non-linear interactions between the zonal flow and MRGWs (Lindzen and Holton 1968; Holton and Lindzen 1972), this suggests that MRGWs should be more active in March-June than in July-February. In addition, Dunkerton (1991a) considered the MRGW seasonal variation at several equatorial stations with records from 1973-87. At Singapore, he showed that the MRGW meridional velocity power [P(v)] maximized during Feb-April at 70-hPa, while 100-hPa and 150-hPa $P(v)$ maximized in the northern summer with a secondary maximum in January. Dunkerton also found that the MRGW intensity did not vary seasonally at all locations, but that a "zonal average" P(v) showed MRGWs in the lower stratosphere at all times of the year with a preference for the northern winter and spring. In addition, he showed that the upper troposphere mean flow in the central Pacific has large annual and semiannual variation with strong easterlies in the northern hemisphere summer, suggesting less MRGW activity in the northern summer.

Maruyama (1991) performed the first long time-period MRGW analysis, examining the

seasonality of MRGWs at Singapore from 1961-89. He found enhancements in 30-hPa $P(v)$ occur slightly after the maximum in the gradually weakening westerly phase of the mean zonal wind. Although this QBO-synchronized oscillation is strong at 30-hPa, it is weak at 50-hPa and 70-hPa. Maruyama also showed that there is a strong annual oscillation in $P(v)$ at 70-hPa (also present but weaker at 30-hPa and 50-hPa) with a maximum around March and a minimum during July-December. This is consistent with the rapid descent of the easterly wind regimes during March-May and slow descent during July-December.

In an extension to his earlier work, Dunkerton (1993) considered stratospheric and tropospheric MRGWs at central and western Pacific stations from 1973-92. He found that the stratospheric MRGWs tend to maximize in January-April. This was also shown to be the time when tropospheric waves were of the same zonal wavenumber as the stratospheric waves (wave number 4). In addition, stratospheric MRGWs maximized in the northern hemisphere winter and spring during the westerly phase of the annual cycle and the westerly phase of the QBO. Finally, the stratospheric MRGWs generally were found not to be coherent with tropospheric oscillations, maximizing at different times of the year. In examining the horizontal structure and propagation of tropospheric MRGWs with both analyzed and rawinsonde data, Dunkerton and Baldwin (1995) concluded that MRGW activity is the strongest in summer, autumn, and winter over the western, central, and eastern Pacific, respectively.

It is clear from the above review that conclusions about the seasonal variability of the upper-tropospheric and lower-stratospheric MRGWs differ by location, height, and possibly by analysis method. Given the extreme convection over the western Pacific in the Northern Hemisphere (NH) summer, it is somewhat surprising that previous seasonal analyses (e.g., Maruyama 1991, Dunkerton 1991a,1993) do not show a summer peak in MRGW activity. Considering that the climatological monthly precipitation over many central and western Pacific stations peaks in the summer (e.g., Terada and Hanzawa 1994) and that MRGWs are likely forced by convection, we hypothesize that there should be a distinct summer peak in central and western Pacific MRGW activity in the lower stratosphere. Thus, the intent of this paper

is to re-examine the average seasonality of lower-stratospheric MRGWs at several stations in the central and western Pacific over a 31-year period (1964-94). Unlike previous studies, we use the seasonally varying spectral analysis (SVSA) method developed by Madden (1986) and a recently developed method in the signal processing literature known as autoregressive cyclic spectral analysis (e.g., Sherman and White 1995). The results of these analyses suggest a twice-yearly peak in lower-stratospheric v-wind power and the mean squared coherence between the u- and v-winds. In turn, this suggests a twice-yearly peak in MRGW activity over the central/western Pacific region. Furthermore, the seasonally varying co-spectrum suggest a variation in equatorial momentum flux convergence associated with the MRGWs, with opposite flux convergences for each semiannual peak. The cyclic spectral analysis also suggests that the fundamental frequency of the maximum v-wind power varies with season. The data and methods are outlined in the next section. Section 3 presents the results of the analyses. A discussion of the results is given in section 4, and a conclusion is presented in section 5.

2 Data and Methods

2.1 Data

Time series of u- and v-components of the wind were extracted from rawinsonde archives at the National Center for Atmospheric Research (NCAR) at four tropical Pacific stations: Koror (7°20′N,134°29′E), Truk (7°27′N,151°50′E), Ponape (6°58′N, 158°13′E), and Majuro (7°05′N,171°23′E) for the years 1964-1994. The geographical locations of these stations is shown in Fig. 1. The period 1966-1972 and 1989-1994 consisted primarily of twice daily observations (00Z and 12Z), while the remaining years consisted of once daily observations (00Z). Except where noted, the analyses presented in this paper use the 00Z observations from the 31-year data record. This 31-year record allows us to examine the long-term climatological behavior of the MRGWs. During quality control, suspect and missing data were replaced by linear interpolation. This study primarily focuses on the 70-hPa level.

2.2 Methods

Spectral and cross-spectral analyses were performed using the Daniell (1946) smoothed periodogram method (e.g., Marple 1987, p.153). A split-cosine bell data taper window (10% on each end) was applied to the data before the periodogram was calculated (e.g., Bloomfield 1976, p.84), although the results were not sensitive to this taper.

To study the seasonal variation of MRGWs over the 31-year data record, the seasonally varying spectral analysis (SVSA) method developed by Madden (1986) and further outlined in Gutzler and Madden (1993) was used. A summary of the method is included in Appendix A. In essence, the technique calculates the seasonal variation of the spectral component associated with a time series after application of an *a priori* selected band-pass filter. In our case, we apply a filter covering the MRGW frequency range to the u- and v-wind time series and then obtain the seasonally varying spectra, as well as the seasonally varying mean-squared coherence (MSC) and the associated phase.

In order to understand the complete frequency-time seasonal variation of power in the MRGW frequency range, a cyclic spectral analysis is performed. Although the periodically correlated nature of atmospheric signals has been known for some time (e.g., Monin 1963; Jones 1964; Jones and Brelsford 1967; Hasselmann and Barnett 1981; Ortiz and Ruiz de Elvira 1985), historically there has not been much of an attempt to use the associated redundancy to improve time series analysis. Recently, there has been a surge of interest in these ideas in the engineering and time series literature, as well as the atmospheric science literature (e.g., Lund et al. 1995; Huang and North 1996). In particular, Huang and North (1996) demonstrated the utility of applying cyclic spectral analysis to cyclostationary atmospheric processes. They used a Discrete Fourier Transform (DFT) implementation of cyclic spectral analysis and applied it to a stochastic climate signal. Following the work of Sherman and White (1995), we take advantage of the high resolution properties of autoregressive (AR) spectral estimators and implement an AR cyclic spectral analysis. We apply this technique to the lower stratospheric wind time series. A description of the methodology is included in Appendix B.

3 Analysis

3.1 Identification of Mixed Rossby-Gravity Waves

Rossby-gravity waves in the lower stratosphere are generally of wave number 4, with a period of 3-5 days and westward phase propagation (e.g., Andrews et al. 1987). Furthermore, the MRGW theoretical structure suggests that the v-component of the wind should lead the u-component by a quarter cycle ($\pi/2$) in the Northern Hemisphere. Thus, we can identify the presence of MRGWs from the spectra of the u- and v-winds, as well as via the mean-squared coherence (MSC) and associated phase between u and v. (MSC is not a useful indicator near the equator as the theoretical u-component of the MRGW is zero there.)

As stated previously, the rawinsonde archive for our analysis stations includes twice-daily observations during two periods, 1966-1972 and 2 October 1989 through 29 November 1994. We extracted twice daily time-series from both periods, each series containing 3770 observations (i.e., 1885 days in each of 10/2/66 - 11/29/71 and 10/2/89 - 11/29/94). Two periods of equal length were chosen to account equally for possible interdecadal biases in MRGW activity. The u- and v-spectra and cross-spectra were calculated for each data period, separately for each station, and the results were averaged (analogous to a Bartlett (1948) smoothing of the Daniell smoothed periodograms). The power spectral estimates, MSC, and phase plots are shown in Fig. 2. Although the u-component power $[P(u)]$ shown in Fig. 2a does not deviate significantly from a red noise spectrum for frequencies below 0.6 cycle/day (cpd), the v-component power $[P(v)]$ shown in Fig. 2b shows a significant peak (relative to a 95% confidence level derived from an AR(1) red noise null hypothesis) in the frequency range between .22 - .38 cpd, corresponding to periods of 2.6 - 4.5 day. There are other marginally significant peaks at .44, .54, .60, and .73 cpd, possibly associated with inertial gravity wave modes. These peaks are not investigated in this paper, but are of concern for their possible aliasing contributions when once-daily observations are used in the seasonal analysis. The MSC between the u and v time series $[MSC(u,v)]$, shown in Fig. 2c, exhibits a distinct peak near .28 cpd (3.6 day). The area of significant MSC extends throughout much of the frequency domain, with minor peaks at .05,

.45, .60, and .73 cpd. The $MSC(u,v)$ phase plot in Fig. 2d shows that the v-wind does indeed lead the u-wind by a quarter of a cycle ($\pi/2$) in the MRGW frequency band. Note that v lags u by a quarter cycle in the .55 - .78 cpd range.

To study the average propagation properties of wave-like disturbances, we perform a cross-spectral analysis between time series at different stations for the frequency band of interest. In particular, we consider the cross-spectral MSC analysis between the v-components at each of the six possible combinations of the four stations, as shown in Fig. 3. If the phase is between zero and $(-\pi)$ then the v-wind at the second station leads that at the first station. The reverse is true if the phase is positive. Thus, considering the longitudinal geographical distribution of the stations, the phase propagation direction can be determined (e.g., Yanai et al. 1968; Dunkerton 1993). For instance, at .25 cpd the Truk-Majuro phase is negative, implying that the Majuro v-wind leads that at Truk. Similarly, the Truk-Koror phase at the same frequency is positive, implying that the Truk v-wind leads that at Koror. Since Koror is west of Truk, and Truk is west of Majuro, then the phase propagation must be towards the west. Furthermore, the magnitudes of these phase differences are indicative of the horizontal scale of the wave. If the phase is plotted as a function of the difference in longitude (at a representative latitude) between the stations, then it is easily shown that the slope of the line is equal to the wavenumber. Based on Fig. 3, a linear regression of the longitude difference versus the $MSC(v,v)$ phase gives a wavenumber (slope) of 5.7 and 2.7 for periods of 5 and 4 days, respectively. Dividing the earth's circumference (at average station latitude 7.2°) by the wavenumber gives respective wavelengths of 7.0×10^3 and $14.8 \times 10^3 km$. These wavelengths imply phase speeds of 16 and 43 ms^{-1}, respectively. Clearly, the smaller phase differences between stations implies lower wavenumbers, larger wavelengths, and faster phase speeds. Figure 3 then suggests that MRGWs with frequencies between .15 - .22 cpd differ dynamically from those in the frequency range .22 - .33 cpd. In addition, there is a distinct phase transition at .33 cpd, suggesting that the MRGW is no longer an appropriate characterization at frequencies above .33 cpd.

Considering the region of significant $P(v)$ shown in Fig. 2b, along with the phase structure transitions shown in Fig. 3, we will focus our seasonal analysis on the frequency band between .22 - .33 cpd. It should be noted that not all of the power $P(v)$ in this frequency band is associated with MRGWs. In fact, the MSC analysis in Fig. 2c suggests that less than 40 percent of the variance in this band may be characterized as such. By focusing on this frequency band, we are simply improving the likelihood that we are considering MRGWs.

3.2 Seasonal Variation of MRGWs

The SVSA methodology described in Appendix A was applied to the once-daily 70hPa Koror time series over the 31 years from 1964-1994. As stated in the previous section, we will focus our seasonally varying (SV) analysis in the frequency band between .22 - .33 cpd. The sixth-order Butterworth bandpass filter (zero phase change) used to extract this frequency band from each time series for the SVSA analysis is shown in Fig. 4 (solid line). Since we are using once-daily observations, we must be concerned about possible aliasing of spectral power in the frequency region between 0.5 - 1.0 cpd. The minor peaks in $P(v)$ shown in Fig. 2b at .54 and .60 cpd do not alias into our MRGW frequency band, but the .73 cpd peak does. The power associated with this peak, however, is minimal compared to that within the MRGW frequency band. Thus, we do not expect our results to be overly contaminated by aliased power from higher frequency waves.

The seasonal variance estimates of the filtered u- and v-winds at Koror are shown in Fig. 5, along with the SV $MSC(u,v)$ and phase. The SV estimates were smoothed by a low-pass Butterworth filter of order 3 and with half-power period of 30 days (Fig. 4, dashed line). The SV MRGW u-wind in Fig. 5a shows peaks in December-January and July, however, we do not focus on the u-wind since the $P(u)$ in the MRGW frequency range is not significant (Fig. 2a). The SV $P(v)$, shown in Fig. 5b, also exhibits semiannual peaks, but in mid January - February and mid August - October. The v-wind peaks are significant when compared to the 95th percentile of 500 identical SVSA analyses of simulated time series with cross-spectral structure similar to the Koror u- and v-winds, but forced by Gaussian random noise (which

should, on average, show no seasonal preference). These simulations were performed assuming a stochastic bivariate AR(4) process. The details of this simulation are included in Appendix C. It is important to reiterate that not all of the seasonally varying power in the MRGW frequency band is associated with MRGWs. Thus, it is beneficial to consider the MSC analysis as well. The SV MSC analysis (Fig. 5c) also shows a semi-annual signal with maxima in mid January - March and mid June - September. These peaks are also significant at the 95% level based on the SVSA simulations. The phase shown in Fig. 5d exhibits variability about the theoretical value ($-\pi/2$), with the first semiannual peak corresponding to a more negative phase (v leads u by more than a quarter cycle), and the second peak corresponding to a less negative phase (v leads u by less than a quarter cycle). The implications of this phase structure will be discussed in below.

In addition to the SVSA v-wind and MSC results for Koror, the results from the other three stations are shown in Fig. 6. The SV v-wind signal does not show a distinct semiannual signal at Truk (Fig. 6b), but does show a strong semiannual signal at Ponape (Fig. 6c) and Majuro (Fig. 6d). Both of the western-most stations (Ponape and Majuro) show a winter maximum and a summer maximum, although there is a phase difference between these stations and Koror. In fact, the Majuro $P(v)$ maxima are nearly a quarter cycle out of phase with those at Koror. One possible explanation for this phase differential between the eastern-most and western-most stations is that the seasonality of the MRGWs could be linked to the semiannual cycle in the global divergent circulation. In particular, Chen and Wu (1992) show that the semiannual cycle in the tropical divergent circulation has two centers, located at approximately 120°E and 160°W, which generally have opposite phase. Thus, it is possible that Majuro and Ponape are influenced by the eastern divergent center, and Koror is influenced by the western center. The MSC plots (Figs. 6e-h) corroborate this, at least with regards to the winter maximum. In general, the MSC plots show twice-yearly peaks, one occurring in the winter-early spring and the other in the summer-early fall [with the possible exception of Ponape (Figure 6g), which only shows a distinct summer peak, and Truk (Figure 6f), which shows a peak in the spring

rather than winter-early spring].

To summarize the SV analysis, we claim there is evidence that suggests MRGWs occur at all times of the year, but they show increased activity (i.e., the waves are either stronger or occur more frequently) twice a year. In general, one peak occurs in winter - early spring and the other peak occurs in the summer - early fall. Although qualitatively similar in most cases, there are noticeable differences between the SV v-wind and SV MSC analyses. The SV MSC analysis (Fig. 5c) suggests that, at most, 23 percent of the SV variance in the v-wind can be attributed to MRGWs. It may be that the v-wind in the MRGW frequency band is contaminated by other tropical disturbances, midlatitude disturbances which have propagated into the equatorial region (e.g., Magaña and Yanai 1995), or "noise". The MSC and phase results between u and v should not be as sensitive to such contamination since the simultaneous analysis of multichannel data acts to filter non-coherent processes. Even so, the signal should be (and typically is) evident in both the v-wind and MSC analyses.

3.2.1 Implications of Seasonally Varying Phase

It was noted in the previous section that there is seasonal variation about the theoretical MRGW phase and that this variability is opposite for the two peaks of the semiannual cycle evident in the MSC plot (Fig. 5c). Figure 7a shows in detail the seasonally varying phase for Koror. The variability in phase about the theoretical value of $(-\pi/2)$ suggests that there may be some dynamical differences between the MRGWs occurring during the first peak as compared to those during the second peak.

From the definition of the $MSC(u,v)$ phase angle (A.6) it is clear that a theoretical phase angle of $-\pi/2$ (i.e., v leads u by 90°) implies that the SV "cospectrum", given in Appendix A by (A.3), must be zero (and is not zero when the phase deviates from this theoretical value). The SV "cospectrum", however, is simply the SV covariance between the filtered u- and v-winds and thus, is a measure of horizontal eddy momentum flux (u^*v^*, where the $*$ denotes deviation from the zonal mean). The SV "cospectrum" for Koror at 70-hPa is shown in Fig. 7b. Clearly, the "cospectrum" variability about zero follows the phase variability about $-\pi/2$.

Thus, the phase difference between the SV MSC semiannual peaks (February and August) implies a "cospectrum" sign difference, and necessarily a difference in horizontal momentum flux for the two semiannual maxima.

To illustrate the effect of the differences in phase/momentum flux, we make use of the theoretical beta-plane MRGW solutions (e.g., Andrews et al. 1987, p.205). In this case, the u- and v-winds associated with the horizontal MRGW are given by:

$$u = v_0 u'(y) e^{i(kx-\omega t)} \quad (1)$$

$$v = v_0 v'(y) e^{i(kx-\omega t-\gamma)} \quad (2)$$

where

$$u'(y) = \frac{i\beta y(1+\frac{k\omega}{\beta})}{\omega} exp[\frac{-(1+\frac{k\omega}{\beta})\beta^2 y^2}{2\omega^2}] \quad (3)$$

$$v'(y) = exp[\frac{-(1+\frac{k\omega}{\beta})\beta^2 y^2}{2\omega^2}] \quad (4)$$

and ω is the MRGW frequency, k is the zonal wave number, γ is a phase shift from the theoretical MRGW v-wind (which we have added for illustration), v_0 is the v-wind amplitude at the equator, $\beta = (2\Omega\cos\phi)/a$, Ω is the earth's angular speed of rotation, ϕ is the latitude, and a is the mean radius of the earth. Figure 8a shows the 4-day, wavenumber 4 theoretical NH MRGW wind-field where v leads u by 90° (i.e., $\gamma = 0$). Contours of horizontal eddy momentum flux (u^*v^*) are shown as well. The meridional derivative of the zonally average momentum flux ($d[u^*v^*]/dy$, where the brackets represent a zonal average) is shown in Fig. 8b to be zero at all latitudes. We next allow the MRGW v-wind to lead the u-wind by less than 90° using the γ parameter in (2). Figure 8c shows the theoretical MRGW wind field and horizontal eddy momentum flux when v leads u by only 78.5°, corresponding to the August semiannual peak at Koror. Figure 8d indicates that there is divergence of horizontal momentum flux from the equator to 4°N and convergence from 4°N to 15°N. If the v-wind leads the u-wind by 101.5° (corresponding to the February semiannual peak at Koror), then the plot of the convergence of momentum flux is symmetrically opposite (not shown) to that in Fig. 8d, implying convergence of horizontal momentum flux near the equator during the

February semiannual peak, and divergence north of 4°N. Thus, the MRGW dynamical structure appears to be fundamentally different during the February and August semiannual peaks. We note that Dunkerton and Baldwin (1995) found the horizontal momentum flux to be higher in the westward propagating tropical depression disturbances than for tropospheric MRGWs. Thus, it is possible that such disturbances are contaminating our analysis, but not likely in our case since we are concerned with lower stratospheric waves, which should not be significantly influenced by the tropospheric tropical depressions.

3.3 Cyclic Spectral Analysis

The seasonally varying analysis of the previous section suggested that there are significant twice-yearly peaks in MRGW activity, and that the characteristics of the MRGWs may be different at different times of the year. The winter and summer peaks in MRGW activity agree with the modeling study of Hayashi and Golder (1980) and the upper tropospheric results of Dunkerton (1991a), but appear to be at odds with the stratospheric work presented by Dunkerton (1991a, 1993) and Maruyama (1991). They did not find a summer- early fall peak in lower stratospheric MRGW activity in the equatorial Pacific lower stratosphere. Thus, to test independently for the presense of the summer maximum in $P(v)$ and to examine the possibility of a seasonal shift in the frequency of the maximum $P(v)$, we employ a cyclic spectral analysis. Such an analysis makes optimal use of the cyclostationary properties of the seasonal atmospheric data and allows us to examine the seasonal variability throughout the complete frequency range of the data. Our implementation is based on an autoregressive (AR) cyclic spectral estimator as described in Appendix B.

Figure 9 shows the v-wind AR(5) cyclic spectra for the portion of the frequency domain of interest in this study (.15 - .36 cpd). The spectra have been smoothed in time with a zero phase-shift Butterworth low-pass filter of order 3 and with a half-power period of 60 days. Note that if we average over the frequency band used in the SVSA in the previous section (.22 - .33 cpd) then the result (not shown) is qualitatively similar to the SV $P(v)$ plot shown in Fig. 5b, with the possible exception of the late December - early January period. The most

notable features of Fig. 9 are the winter maximum, the clear minimum in spring, the summer maximum, and a less intense minimum in the fall. The winter maximum is generally in the lower frequency range (.22 - .30 cpd) while the summer maximum is in the higher frequencies (.28 - .35 cpd). In particular, it seems that the maximum in $P(v)$ initiates in July at .35 cpd and slowly migrates to lower frequencies as the year progresses. Although not all of the v-wind power is associated with MRGWs, these results suggest that the background zonal wind may be modifying the frequency characteristics of the MRGWs. Figure 10 shows the average zonal wind for Koror (solid line). Clearly, there is a strong annual cycle component to the zonal wind, with maximum easterlies in August and minimum easterlies in February. The minimum in $P(v)$ occurring in the spring corresponds to the period of rapidly increasing easterlies in the zonal wind, while the peaks correspond to the period of decreasing easterlies. This is consistent with Maruyama's (1969) finding that MRGWs are more prevalent when the absolute value of the zonal wind speed is decreasing in time.

Figure 11 shows contours of the AR(5) cyclic spectrum for Majuro at 70hPa. Note that the semiannual cycle in $P(v)$ is more distinct than for the Koror case, primarily because the spring and fall minima are better defined, even though the August peak is less intense than at Koror. The average Majuro zonal wind (Fig. 10, dotted line) shows a strong seasonal cycle with a more pronounced peak in May and a less intense minimum in August (relative to Koror). The Majuro cyclic spectra plot shows that the frequency of the maximum in $P(v)$ varies with season, but the gradual frequency shift with time shown at Koror is not present.

4 Discussion

We have demonstrated by SVSA that there are two seasonal periods of increased $P(v)$ and $MSC(u, v)$ in the western Pacific equatorial lower stratosphere. The $P(v)$ maxima were confirmed by cyclic spectral analysis. Although the presence of the winter - early spring and summer - late fall peaks has been shown in numerical modeling (Hayashi and Golder 1980), the latter peak is at odds with recent observational studies (e.g., Maruyama 1991; Dunkerton

1991a,1993). In addition, assuming that the peaks are associated with MRGWs, the summer maximum seems to contradict the belief that MRGWs should not occur when the mean zonal wind is easterly. We address these issues and discuss the implications of these findings in the following paragraphs.

We first address the apparent contradiction with recent observational results, which suggest a late winter - early spring maximum in MRGW activity. One difference between our study and those presented in Maruyama (1991) and Dunkerton (1991a) is that they focus their seasonal analyses on Singapore (or, in Dunkerton's case, much of the presentation). We note that the Maruyama (1991) and Dunkerton (1991a) results were based on a moving window spectral analysis. Although this technique can give smoothed estimates of the seasonally varying spectral power, the relatively short window length and DFT implementation limits the resolution of this approach and this could be the source of the difference. The SVSA and cyclic spectral approaches used in the present study have superior time resolution when compared to the moving window approach (but are not without their limitations, as discussed in the Appendix B), and thus should be more sensitive to seasonal variability. Therefore, we performed the SVSA on 70mb data from Singapore. The analysis gives results (not shown) very close to that presented in Fig. 3 of Dunkerton (1991a) [i.e., it shows a strong spring peak but only a small July peak in seasonal $P(v)$]. Thus, it does not appear that the differences between our results and previous studies are related to the differences in analysis methods (at least as far as Singapore is concerned).

If we accept that the seasonal peaks in $P(v)$ and $MSC(u,v)$ are related to MRGWs, then the lack of a summer MRGW maximum at Singapore is due either to MRGW forcing differences between Singapore and the central/western Pacific stations, or to a damping mechanism that suppresses MRGW waves generated in the central/western Pacific as they propagate to the Singapore region during the summer. First, note that Singapore is closer to the equator and much further west than the stations included in our analysis (see Figure 1). Thus, although the MRGW v-wind is stronger near the equator, it is possible that the semiannual signal

is not as strong there as for the western Pacific stations near 7°N that we considered. As stated previously, convection is a possible forcing mechanism for MRGWs, and is certainly linked to MRGW activity (at least in the troposphere). This forcing mechanism should have a strong impact in the summer months, when convection over the western Pacific warm pool is the strongest. Perhaps, the lack of a summer peak in MRGW activity at Singapore is related to less convective activity over that region. In fact, examination of climatological mean monthly precipitation for Singapore (Nieuwolt 1984) and our analysis stations (Terada and Hanzawa 1984) corroborates this speculation. While the four central/western Pacific stations of our analysis show significant summer precipitation, Singapore shows a climatological minimum in the summer. Such a minimum in summer precipitation over the Singapore region is also suggested by the SSM/I-derived monthly rainfall estimates presented by Berg and Avery (1994). This summer minimum in convection is most likely attributable to the influence of the climatological high pressure center located over northwest Australia, which extends into the Indonesian region, effectively splitting the Intertropical Convergence Zone (e.g., see the July plots of SSM/I derived monthly rainfall in Fig. 3 of Berg and Avery 1994).

Dunkerton (1991a) also looked at the seasonally varying $P(v)$ for an average of six stations, and Dunkerton (1993) used many of the same western Pacific stations as we have used, but did not find a strong semiannual signal in either case. Dunkerton (1991a) does not show seasonally varying results for the stations we have considered here. However, he does present results for a 6-station "zonal average" which does not show a summer peak. It is possible that the summer peak is only present in the central/western Pacific and is thus "averaged out" of his analysis. Although Dunkerton (1993) considers many of the same central/western Pacific stations as we have, the 70-hPa time varying analysis he describes is not shown in his paper, so that a direct comparison to our results is not possible. It is implied in that paper that the seasonality was detected from a visual analysis of the time series of the windowed DFT $P(v)$ values. Although there is utility to such an approach (especially for characterizing interannual variability), the inherent subjectivity may lead to an oversight of some climatological features.

We now address the observation that the summer peak in MRGW occurrence coincides with the average maximum easterlies in the zonal wind (e.g., Fig. 10). As Maruyama (1991) notes, his Fig. 5 also shows several instances when $P(v)$ maximizes during the maximum easterly phase of the zonal wind (e.g., 1982). Although this was attributed to possible data errors, a close look at these results suggest that this maximum in $P(v)$ occurs quite frequently during periods of maximum easterlies. Linear theory predicts that MRGWs can only propagate vertically when the zonal flow is westerly (e.g., Lindzen 1970, 1971; Andrews et al. 1987). However, although we would not expect the MRGWs to propagate vertically above 70hPa during the summer maximum, there is no limitation on their existence or excitation at this level and time. In fact, a plot of the SV $P(v)$ at the 30-hPa level for Koror (Fig. 12) shows the winter - early spring maximum, but not the summer maximum. This is consistent with the inability of the MRGWs to propagate vertically in the summer.

If we accept that there are indeed two maxima in MRGW activity, then it appears that the dynamical characteristics of the MRGWs are different for these two peaks. As shown in Section 3.2.1, this is supported by the SV phase difference, which implies a reversal in the convergence of horizontal momentum flux associated with the MRGWs from each maximum. Furthermore, the cyclic spectral analysis results of Section 3.3 show that the frequency is higher for the summer $P(v)$ maximum than for the winter - early spring maximum. Such differences may be simply due to the seasonal variation of the basic state zonal wind. However, these differences could also have implications as to the excitation mechanism for the MRGWs. As mentioned in the Introduction, there is still considerable debate in the atmospheric science community as to whether MRGWs are forced laterally or are coupled to convection. As the observational evidence can support either theory, it seems plausible that both mechanisms are valid. It is then possible that the MRGWs associated with the summer - early fall maximum are mainly due to one type of forcing, while the winter - early spring waves are primarily due to the other forcing mechanism. As suggested in the contrast between summer MRGW activity in Singapore versus the central/western Pacific stations, it is likely that much of the summer

maximum is related to convective forcing. In addition, due to the increase in extra-tropical baroclinic wave activity in the NH winter, it is possible that the winter peak may contain more influence from lateral forcing. However, such speculations should be tested by numerical simulations.

5 Conclusions

We analyzed 31 years of lower stratospheric wind data at four stations in the western/central Pacific. Traditional spectral and cross-spectral analysis led us to conclude that there is a significant signal between 3 - 4.5 days, and that the waves in this range could be characterized as mixed Rossby-gravity waves (MRGWs). We then applied the seasonally varying spectral analysis (SVSA) method developed by Madden (1986) to study the average seasonal variation of these waves. The SVSA suggested that there are significant twice-yearly maxima in v-wind power and the mean-squared coherence between the u- and v-wind, with peaks occurring in winter - early spring and in summer - early fall. These results were confirmed by a cyclic spectral analysis. It was then suggested that these peaks are indicative of increased MRGW activity. In addition, the SVSA mean-squared coherence between the u- and v-winds and the associated phase implied that there is convergence of horizontal momentum flux associated with these waves, and that the sign of that convergence is different during the two maxima. The cyclic spectral analysis also suggested that the frequency of the v-wind power is different during the two maxima. If we assume MRGWs are associated with such peaks, then these differences could be due to either the seasonal variation of the basic zonal state or to different MRGW forcing mechanisms (i.e., convective versus lateral excitation mechanisms).

It was noted that a twice-yearly peak in MRGW activity contradicts some recent observational studies, but is certainly not without precedence. Clearly, further efforts are required to help resolve these contradictions. In particular, it would be beneficial if additional cyclic spectral analyses were conducted with similar data sets. As mentioned in Appendix B, there is ongoing research exploring more sophisticated implementations of the AR cyclic spectral

analysis used in this study. It would also be useful to develop a multi-channel AR cyclic spectral analysis technique, which would allow cyclic cross-spectral analysis of the u- and v-winds. Although this has been implemented with the DFT approaches (e.g., Huang and North 1996), it is an area of current research with AR implementations. In addition, modeling studies should be used to explore the semiannual peaks in MRGW activity, and their links to possible forcing mechanisms. Finally, it is not clear whether the seasonal increases in spectral and cross-spectral intensity observed in this study are due to increased amplitudes of the MRGWs or increases in occurrence. Further analysis is required to resolve this question.

Acknowledgements

The research was sponsored by the U.S. Department of Energy, Office of Energy Research, Environmental Sciences Division, Office of Health and Environmental Research, under the first author's appointment to the Graduate Fellowships for Global Change administered by Oak Ridge Institute for Science and Education. Data and computer time were generously provided by NCAR during summer visits by the first author. Additional support was provided by NSF Grant ATM-9416954. We wish to thank Prof. Peter Sherman for his discussions related to AR cyclic spectral analysis.

Appendix A: Seasonally Varying Spectral Analysis (SVSA)

To study the seasonal variation of MRGWs over the 31 year data record, we used the SVSA procedure outlined in Madden(1986) and Gutzler and Madden (1993). A brief summary of the method follows.

The annual variation in the mean was removed from each time series. This was accomplished by first computing averages over all years for each day of the year. This average series was then subjected to Fourier analysis and a weighted sum of the first five harmonics and the annual mean were removed from each year of data.

Given time series of the zonal, $u(t)$, and meridional, $v(t)$, winds, the data were divided

into 540 day segments, each starting on 1 July of a given year. A sixth-order Butterworth bandpass filter with zero phase shift (e.g., Hamming 1989) covering the frequency range of the phenomenon of interest was applied to each of the 540 day segments, which we denote by $U_f(t)$ and $V_f(t)$ for $u(t)$ and $v(t)$, respectively. Estimates of the seasonally varying wind variances are then given by:

$$S_U^f(t) \equiv \langle U_f(t)^2 \rangle, \tag{A. 1}$$

and

$$S_V^f(t) \equiv \langle V_f(t)^2 \rangle, \tag{A. 2}$$

where the angle brackets indicate an average over all segments for the t-th day of the year. The seasonally varying covariance is then given by:

$$C_{U,V}^f(t) \equiv \langle U_f(t) \cdot V_f(t) \rangle. \tag{A. 3}$$

The quadrature variance is calculated by shifting one of the series (in our case the $U_f(t)$ series) by one-quarter cycle through a Hilbert transform:

$$Q_{U,V}^f \equiv \langle H[U_f(t)] \cdot V_f(t) \rangle, \tag{A. 4}$$

where H is the Hilbert transform operator (e.g., Barnett 1983). The estimates of $S_U^f(t)$, $S_V^f(t)$, and $C_{U,V}^f(t)$ are then smoothed via a low-pass filter. Estimates of seasonally varying mean-squared coherence and phase can then be estimated from these smoothed estimates by:

$$MSC_{U,V}^f(t) \equiv \frac{\tilde{C}_{U,V}^f(t)^2 + \tilde{Q}_{U,V}^f(t)^2}{\tilde{S}_U^f(t) \cdot \tilde{S}_V^f(t)}, \tag{A. 5}$$

and

$$\Phi_{U,V}^f(t) \equiv \arctan(\tilde{Q}_{U,V}^f(t) / \tilde{C}_{U,V}^f(t)), \tag{A. 6}$$

respectively, where the tilde indicates a smoothed value.

Appendix B: Autoregressive Cyclic Spectral Analysis

Traditional spectral analysis techniques require an assumption of second-order stationarity (i.e., constant mean with autocorrelation depending only on time lag). This assumption clearly

breaks down when the physical process under consideration has known cycles (e.g., solar influenced annual and semiannual cycles in atmospheric processes). In that case the mean and variance are also periodic. Traditionally, investigators remove these cycles, hoping that they then can satisfy the stationarity assumption (which typically, they can't, at least with regard to the variance). However, from a statistical perspective, it makes sense to use the redundant information contained in the periodically correlated moments optimally, rather than to remove it. An excellent discussion of the analysis of periodically correlated atmospheric time series can be found in Lund et al. (1995). In addition, Huang and North (1996) provide a comprehensive discussion of cyclic spectral analysis related to atmospheric processes. A more general and complete discussion of periodically correlated time series analysis from an engineering perspective can be found in Gardner (1994) and references therein. Our discussion of autoregressive cyclic spectral analysis follows that found in Sherman and White (1995), who apply the technique to rotating machinery.

A random process Y_t, for $t \in 1, \ldots, n$ is defined as wide-sense cyclostationary (wsc) if it has an autocorrelation function $E(Y_s Y_t) \equiv R(s,t)$ that is d-periodic:

$$R(s,t) = R(s+md, t+md) \tag{B. 1}$$

for $m \in 0, \pm 1, \pm 2, \ldots$.

Now, consider a d-periodic autoregressive (AR) process:

$$Y_t - \sum_{j=1}^{p_t} a_t(j) Y_{t-j} = \epsilon_t \tag{B. 2}$$

where $p_t = p_{t+d}$ is the AR model order for time t, ϵ_t is a wsc white noise process such that $E[\epsilon(t)] = 0$, $E[\epsilon^2(t)] = \sigma_t^2 = \sigma_{t+d}^2$, and $a_t(j) = a_{t+d}(j)$ are the cyclostationary AR parameters.

The Yule-Walker type equations corresponding to (B.2) are

$$R(t, t-\tau) = \sum_{k=1}^{p_t} a_t(k) R(t-k, t-\tau) + \delta_\tau \sigma_t^2, \tag{B. 3}$$

where $t = 1, \ldots, d$; $\tau = 0, \ldots, p_t$; and $\delta_\tau = 1$ for $\tau = 0$ and zero elsewhere.

Estimation

Pagano (1978) showed that a statistically consistent estimator for the autocorrelation of a one-dimensional wsc process is:

$$\hat{R}(s,t) = N^{-1} \sum_{k=0}^{N-1} Y_{s+kd} Y_{t+kd} \tag{B. 4}$$

where, $N = n/d$; $Y_h = 0$ for $h \leq 0$, and without loss of generality, Y_t is taken to have zero mean. By utilizing the consistent estimators given by (B.4) in (B.3), Pagano (1978) showed that the Yule-Walker equations can be solved, giving consistent estimators of the parameters in the model (B.2). In practice, a common approach is to solve the Yule-Walker set via least squares (Pagano 1978). In addition, a choice must be made for the model orders at each time, p_t. A simple approach is to find the AR model order most appropriate for the time series (e.g., using the Akaike Information Criterion; e.g., Marple 1987, p.229) after having filtered for the frequency band of interest. In our case, we assume that $p_t = p$ for all t, where p is this "average" model order. Such an approach is somewhat naive, but does provide a simple first approximation to the cyclic spectra. The optimal choice of AR model orders is a subject of current research (e.g., McLeod 1994).

Given the autoregressive parameters $a_t(j), t = 1, \ldots, d; j = 1, \ldots, p_t$, we can then use the AR spectral estimator (e.g., Marple 1987; Wikle et al. 1995) applied at each t:

$$AR_{p_t}(\omega) = [|\phi_{p_t}(e^{i\omega})|^2]^{-1} \tag{B. 5}$$

where

$$\phi_{p_t}(e^{i\omega}) = \sigma_t^{-1} \sum_{k=0}^{p_t} a_t(k) e^{-ik\omega}, \tag{B. 6}$$

and $a_t(0) \equiv 1$. In practice, for each t we pad the autoregressive spectral estimates with zeros to increase the resolution of the discrete Fourier transform (DFT) in (B.6) [i.e., we apply the DFT to $\{\hat{a}_t(0), \hat{a}_t(1), \ldots, \hat{a}_t(p_t), 0, \ldots, 0\}$ where $\hat{a}_t(k), k = 1, \ldots, p_t$ are the AR parameter estimates obtained from the solution of the Yule-Walker equations].

We must recognize the limitations inherent in the cyclic spectral analysis (and the SVSA outlined in Appendix A) which have led us to conclude there are two peaks in the v-wind

power. In particular, the cyclic spectral approach assumes explicitly (and the SVSA approach assumes implicitly), that there is an underlying annual cycle which undergoes no variation in period. In fact, this may not be true (especially with daily data). It has been shown (e.g., Sherman and White 1995) that random variation in the cyclostationary period can seriously affect a seasonally varying analysis. Methods have been developed to track the underlying tonal frequency (e.g. using an extended Kalman filter). Atmospheric scientists, however, tend to be reluctant to deviate from the hypothesis of a rigid annual cycle since the year to year solar cycle variability is negligible. Furthermore, Wikle et al. (1995) showed that the 70-hPa annual cycle at a western Pacific station (Truk) has a very strong periodicity, nearly indistinguishable from a sinusoid. Thus, our assumption of a constant cyclostationary period is reasonable in this case. Of course, the windowed DFT method utilized in previous studies (e.g., Maruyama 1991; Dunkerton 1991a) would also suffer from such a misspecification in cycle period.

We also note a limitation in our AR implementation of cyclic spectral analysis. This concerns the choice of model order. More specifically, the concern is related to the assumption of a constant AR model order over time, which, as noted previously, simplifies our analysis. The optimal choice of these time varying model orders is a current area of research in the signal processing community. In the present study, since our results are corroborated by the SVSA method, we believe the constant AR model order assumption does not affect our results critically. However, a more detailed cyclostationary analysis is warranted. Such an analysis should consider an AR implementation with time varying model orders (e.g., McLeod 1994).

Appendix C: Stochastic 2-Channel AR Simulation

Consider a 2-channel autoregressive process:

$$\mathbf{y}_t = \sum_{k=1}^{p} \mathbf{A}_k \mathbf{y}_{t-k} + \mathbf{e}_t, \tag{C. 1}$$

where p is the 2-channel AR model order, $\mathbf{y}_t \equiv [y_{1t}, y_{2t}]'$, $\mathbf{A}_k$ is a 2×2 matrix of AR coefficients, and $\mathbf{e}_t \equiv [e_{1t}, e_{2t}]'$ is a zero-mean wide-sense stationary error process such that

$$E[\mathbf{e}_t \mathbf{e}_t'] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}, \qquad \text{(C. 2)}$$

where $\sigma_{12} = \sigma_{21}$.

We now present an algorithm which allows us to simulate from a 2-channel (e.g., u- and v-wind) AR process of order p.

- Let $\mathbf{y}_t = [u_t, v_t]'$.
- Given $\mathbf{y}_t$ and a model order p (e.g., chosen via the multi-channel Akaike Information Criterion; Marple 1987, p.409), estimate $\hat{\mathbf{A}}_k$, $\hat{\sigma}_1^2, \hat{\sigma}_2^2$, and $\hat{\sigma}_{12}$ by utilizing a method such as the multichannel Levinson recursion algorithm (e.g., Marple 1987, p.400-402).
- Simulate e_{1t} from $N(0, \sigma_1^2)$ [i.e., a normal distribution with mean 0 and variance σ_1^2].
- Simulate ν_t from $N(0,1)$ [i.e., a standard normal distribution].
- Let $e_{2t} = a^2 \hat{\sigma}_1^2 + b^2$, where $a = \hat{\sigma}_{12}/\hat{\sigma}_1^2$ and $b = \hat{\sigma}_2^2 - \hat{\sigma}_{12}^2/\hat{\sigma}_1^2$ can be shown to give the appropriate 2-channel AR error structure when $\mathrm{E}(e_{1t}\nu_t) = 0, \forall t$.
- Use the simulated e_{1t}, e_{2t}, and estimated $\hat{\mathbf{A}}_k, k = 1, \ldots, p$ in (C1) to generate as large a simulation of $\mathbf{y}_t$ as necessary, remembering to let the simulation "burn in" for a reasonable amount of time.
- Check the 2-channel spectra, cross-spectra, mean-squared coherence, and phase to verify that the simulated AR process is realistic in the portion of the frequency domain of interest. If not, the AR model order should be adjusted, and the simulation repeated.

References

Andrews, D.G., Holton, J.R., and C.B. Leovy, 1987: *Middle Atmospheric Dynamics,* Academic Press, 489 pp.

Barnett, T.P., 1983: Interaction of the monsoon and Pacific trade wind system at interannual time scales. Part I: The equatorial zone. *Mon. Wea. Rev.*,, **111**, 756-773.

Bartlett, M.S., 1948: Smoothing periodograms from time series with continuous spectra. *Nature*, **161**, 686-687.

Berg, W., and S.K. Avery, 1994: Rainfall variability over the tropical Pacific from July 1987 through December 1991 as inferred via monthly estimates from SSM/I. *J. Appl. Meteor.*, **33**, 1468-1485.

Bloomfield, P., 1976: *Fourier analysis of time series*, John Wiley and Sons, 258 pp.

Chen, T.-C., and K.D. Wu, 1992: Semi-annual oscillation of the global divergent circulation. *Tellus*, **44A**, 357-365.

Daniell, P.J., 1946: On the theoretical specification and sampling properties of autocorrelated time-series. *J. R. Stat. Soc.*, Ser. B, **8**, 88-90.

Dunkerton, T.J., 1991a: Intensity variation and coherence of 3-6 day equatorial waves. *Geophys. Res. Lett.*, **18**, 1469-1472.

Dunkerton, T.J., 1991b: Nonlinear propagation of zonal winds in an atmosphere with Newtonian cooling and equatorial wavedriving. *J. Atmos. Sci.*, **48**, 236-263.

Dunkerton, T.J., 1993: Observation of 3-6-day meridional wind oscillations over the tropical Pacific, 1973-1992: Vertical structure and interannual variability. *J. Atmos. Sci.*, **50**, 3292-3307.

Dunkerton, T.J., and M.P. Baldwin, 1995: Observation of 3-6 day meridional wind oscillations over the tropical Pacific, 1973-1992: Horizontal structure and propagation. *J. Atmos. Sci.*, **52**, 1585-1601.

Emmanuel, K., 1993: The effect of convective response times on WISHE models. *J. Atmos. Sci.*, **50**, 1763-1775.

Gardner, W.A., 1994: An introduction to cyclostationary signals. *Cyclostationarity in Communications and Signal Processing*. W.A. Gardner, Ed., IEEE Press, 504 pp.

Goswami, P., and B.N. Goswami, 1991: Modification of n=0 equatorial waves due to interaction between convection and dynamics. *J. Atmos. Sci.*, **48**, 2231-2244.

Gutzler, D.S., and R.A. Madden, 1993: Seasonal variations of the 40-50-day oscillation in atmospheric angular momentum. *J. Atmos. Sci.*, **50**, 850-860.

Hamming, R.W., 1989: *Digital Filters*, Third Edition, Prentice-Hall.

Hasselmann, K., and T. P. Barnett, 1981: Techniques of linear prediction for systems with periodic statistics. *J. Atmos. Sci.*, **38**, 2275- 2283.

Hayashi, Y., 1970: A theory of large-scale equatorial waves generated by condensation heat and accelerating the zonal wind. *J. Meteor. Soc. Japan,* **48**, 140-160.

Hayashi, Y., and D.G. Golder, 1978: The generation of equatorial transient planetary waves: Control experiments with a GFDL general circulation model. *J. Atmos. Sci.*, **35**, 2068-2082.

Hayashi, Y., and D.G. Golder, 1980: The seasonal variation of tropical transient planetary waves appearing in a GFDL general circulation model. *J. Atmos. Sci.*, **37**, 705-716.

Hendon, H.H., and B. Liebmann, 1991: The structure and annual variation of antisymmetric fluctuations of tropical convection and their association with Rossby-gravity waves. *J. Atmos. Sci.*, **48**, 2127-2140.

Holton, J.R., 1972: Waves in the equatorial stratosphere generated by tropospheric heat sources. *J. Atmos. Sci.*, **29**, 368-375.

Holton, J.R., and R.S. Lindzen, 1972: An updated theory for the quasi-biennial cycle of the tropical stratosphere. *J. Atmos. Sci.*, **29**, 1076-1080.

Huang, J.-P., and G.R. North, 1996: Cyclic spectral analysis of fluctuations in a GCM simulation. *J. Atmos. Sci.*, **53**, 370-379.

Itoh, H., and M. Ghil, 1988: The generation mechanism of mixed Rossby-gravity waves in the equatorial troposphere. *J. Atmos. Sci.*, **45**, 585-604.

Jones, R.H., 1964: Spectral analysis and linear prediction of meteorological time series. *J. Appl. Meteor.*, **3**, 45-52.

Jones, R.H., and W.M. Brelsford, 1967: Time series with periodic structure. *Biometrika,* **54**, 403-407.

Lund, R.B., H.L. Hurd, P. Bloomfield, and R. Smith, 1995: Climatological time series with periodic correlation. *J. Climate,* **8**, 2787-2809.

Lindzen, R.S., 1970: Internal equatorial planetary-scale waves in shear flow. *J. Atmos. Sci.*, **27**, 394-407.

Lindzen, R.S., 1971: Equatorial planetary waves in shear: Part I. *J. Atmos. Sci.*, **28**, 609-622.

Lindzen, R.S., and J.R. Holton, 1968: A theory of the quasi-biennial oscillation. *J. Atmos. Sci.*, **25**, 1095-1107.

Lindzen, R.S., and T. Matsuno, 1968: On the nature of large-scale wave disturbances in the equatorial lower stratosphere. *J. Meteor. Soc. Japan,* **46**, 215-221.

McLeod, A.I., 1994: Diagnostic checking of periodic autoregression models with application. *J. Time Ser. Anal.,* **15**, 221-233.

Madden, R.A., 1986: Seasonal variations of the 40-50 day oscillation in the tropics. *J. Atmos. Sci.*, **43**, 3138-3158.

Magaña, V., and M. Yanai, 1995: Mixed Rossby-gravity waves triggered by lateral forcing. *J. Atmos. Sci.,* **52**, 1473-1486.

Mak, M.-K., 1969: Laterally driven stochastic motions in the tropics. *J. Atmos. Sci.,* **26**, 41-64.

Marple, S.L., Jr., 1987: *Digital Spectral Analysis with Applications,* Prentice-Hall, 492 pp.

Maruyama, T., 1969: Long-term behavior of Kelvin waves and mixed Rossby-gravity waves. *J. Meteor. Soc. Japan,* **47**, 245-254.

Maruyama, T., 1991: Annual variations and QBO-synchronized variations of the equatorial wave intensity in the lower stratosphere at Singapore during 1961-1989. *J. Meteor. Soc. Japan,* **69**, 219-232.

Matsuno, T., 1966: Quasi-geostrophic motions in the equatorial area. *J. Meteor. Soc. Japan,* **41**, 25–42.

Monin, A.S., 1963: Stationary and periodic time series in the general circulation of the atmosphere. *Proc. Symp. Time Series Analysis,* M. Rosenblatt, Ed., Wiley, 144-151.

Nieuwolt, S., 1984: The climates of continental southeast Asia. In *World Survey of Climate,* Vol. 9, Takahashi, K., and H. Arakawa, Ed., Elsevier, 333pp.

Nitta, T., 1970: On the role of transient eddies in the tropical troposphere. *J. Meteor. Soc. Japan,* **48**, 348-359.

Ortiz, M.J., and A. Ruiz de Elvira, 1985: A cyclo-stationary model of sea surface temperatures in the Pacific ocean. *Tellus,* **37A**, 14-23.

Pagano, M., 1978: On periodic and multiple autoregressions. *Ann. of Statist.,* **6**, 1310-1317.

Sherman, P.J., and L.B. White, 1995: Improved periodic spectral analysis with application to diesel vibration data. *J. Acoust. Soc. Amer.*, **98**, 3285-3301.

Terada, K., and M. Hanzawa, 1984: Climate of the north Pacific ocean. In *World Survey of Climate,* Vol. 15, H. van Loon, Ed., Elsevier, 716 pp.

Wikle, C.K., P.J. Sherman, and T.-C. Chen, 1995: Identifying periodic components in atmospheric data using a family of minimum variance spectral estimators. *J. Climate*, **8**, 2352-2363.

Yanai, M., and Y. Hayashi, 1969: Large-scale equatorial waves penetrating from the upper troposphere into the lower stratosphere. *J. Meteor. Soc. Japan*, **47**, 167-182.

Yanai, M., and M.-M. Lu, 1983: Equatorially trapped waves at the 200mb level and their association with meridional convergence of wave energy flux. *J. Atmos. Sci.*, **40**, 2785-2803.

Yanai, M., and T. Maruyama, 1966: Stratospheric wave disturbances propagating over the equatorial Pacific. *J. Meteor. Soc. Japan*, **44**, 291-294.

Yanai, M., T. Maruyama, T. Nitta, and Y. Hayashi, 1968: Power spectra of large-scale disturbances over the tropical Pacific. *J. Meteor. Soc. Japan*, **46**, 308-323.

Zangvil, A., and M. Yanai, 1980: Upper tropospheric waves in the tropics. Part I: Dynamical analysis in the wavenumber-frequency domain. *J. Atmos. Sci.*, **37**, 283-298.

Zangvil, A., and M. Yanai, 1981: Upper tropospheric waves in the tropics. Part II: Association with clouds in the wavenumber-frequency domain. *J. Atmos. Sci.*, **38**, 939-953.

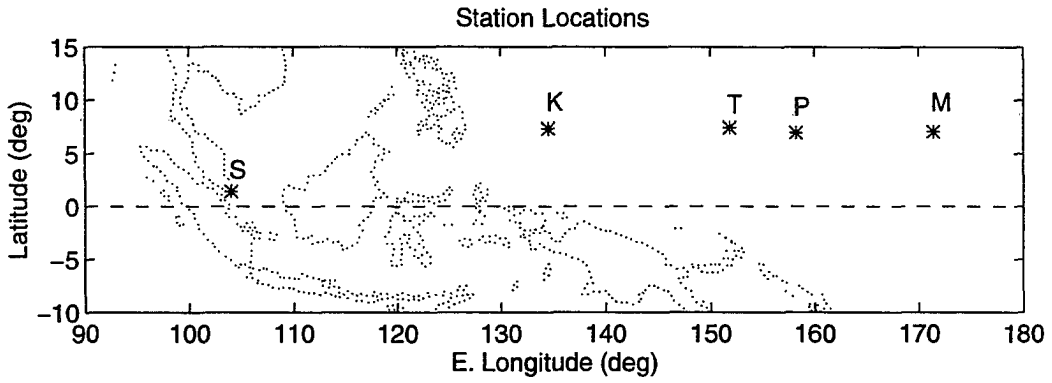

Figure 1 Geographical locations of stations used in the analysis. S-Singapore, K-Koror, T-Truk, P-Ponape, M-Majuro.

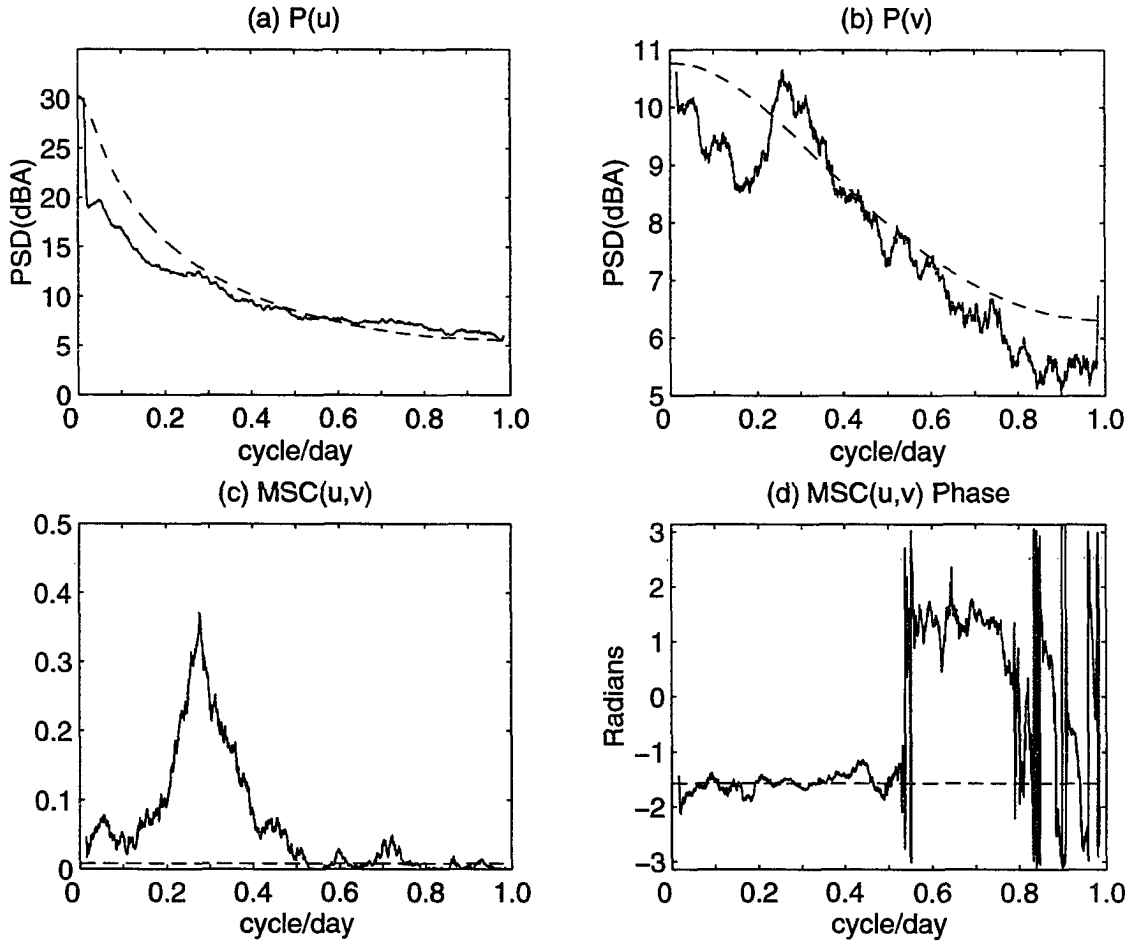

Figure 2 Cross-spectral analysis of twice-daily 70-hPa u- and v-winds at Koror using a composite of two 1885 day periods, 10/2/66 - 11/29/71 and 10/2/89 - 11/29/94. Results were obtained using a smoothed periodogram method with smoothing bandwidth of .028 cycles per day (cpd). Dashed lines represent the 95% confidence level of a red noise [i.e., AR(1)] null-hypothesis. Solid lines represent the: (a) power of the u-wind in decibels (dBA: $dBA = 10log_{10}x$, where x is the power in $(ms^{-1})^2 \cdot day$), (b) v-wind power (dBA), (c) mean-squared coherence (MSC) between the u- and v-wind, and (d) the phase between u- and v-wind in radians.

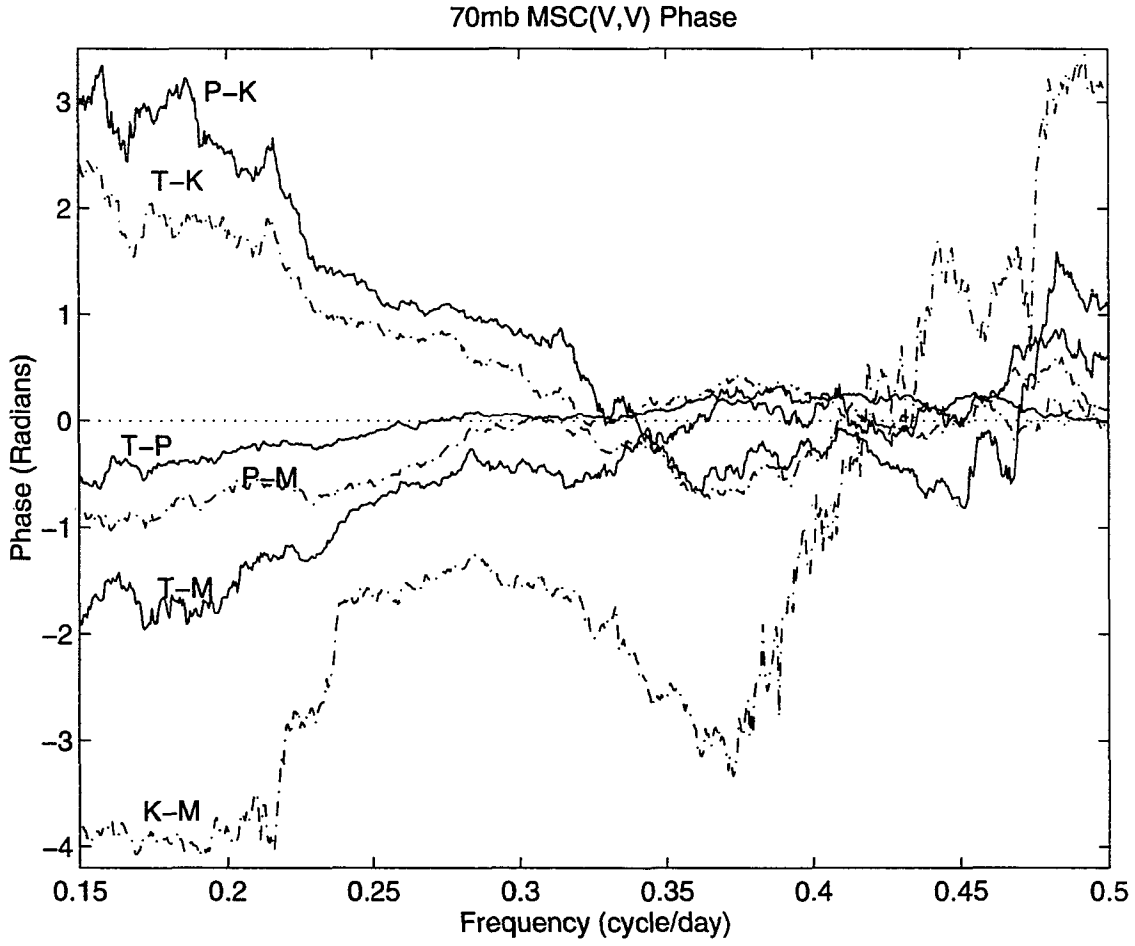

Figure 3 MSC phase (radians) between the 70-hPa v-wind at two stations using data and methods as for Figure 2 except with a smoothing bandwidth of .056 cpd. Note that only a portion of the frequency domain is shown; the MSC (not shown) associated with the phase in this frequency range is relatively large, although decreasing with separation distance. Station symbols are: P(Ponape); K(Koror); T(Truk); M(Majuro), so that, e.g., P-K represents the phase between v at Ponape and v and Koror.

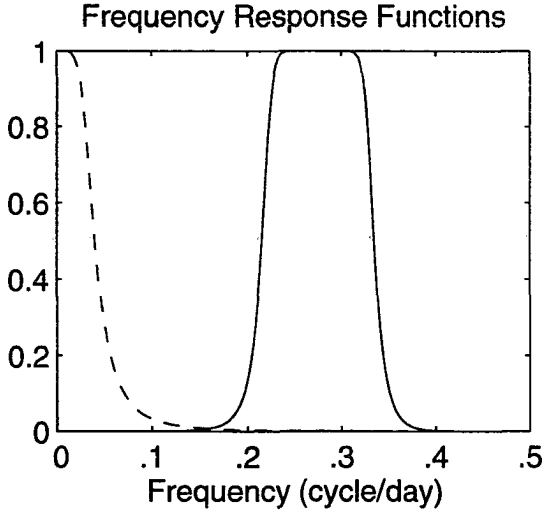

Figure 4 Filter frequency response functions used in the seasonally varying spectral analysis. Solid line: Butterworth bandpass filter (order 6) with half-power points at .22 cpd (4.5 day) and .33 cpd (3 day). Dashed line: Butterworth lowpass filter (order 3) with half-power point at .033 cpd (30 days).

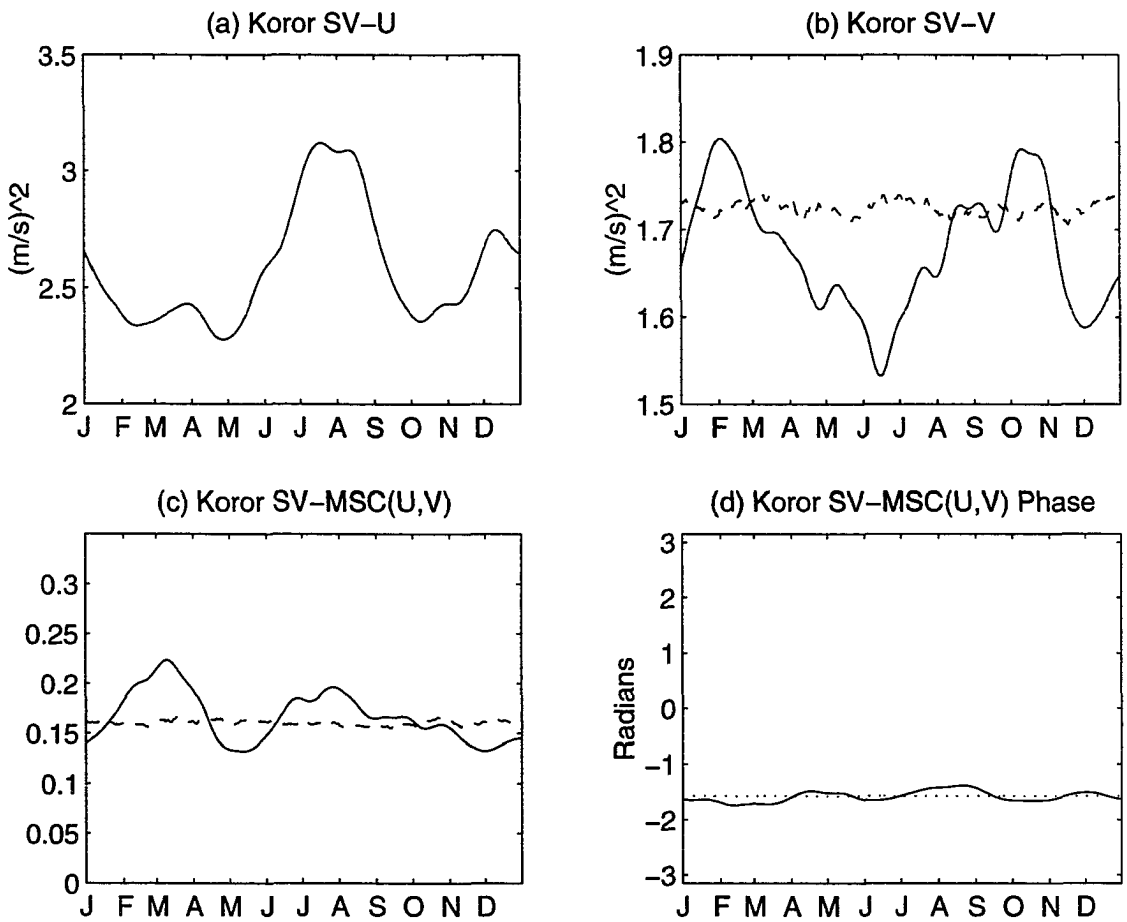

Figure 5 Seasonally varying spectral analysis for 70-hPa Koror u- v-winds with power in the range 3 - 4.5 day. Solid lines: (a) Seasonally varying (SV) u-wind variance $(ms^{-1})^2$, (b) SV v-wind variance $(ms^{-1})^2$, (c) MSC between u- and v-wind, (d) MSC phase (radians). Dashed lines: 95^{th}-percentile significance level estimate obtained from 500 independent SV analyses using simulated data from a 2-channel autoregressive model with similar spectral structure to the observed u- and v-winds (see Appendix C).

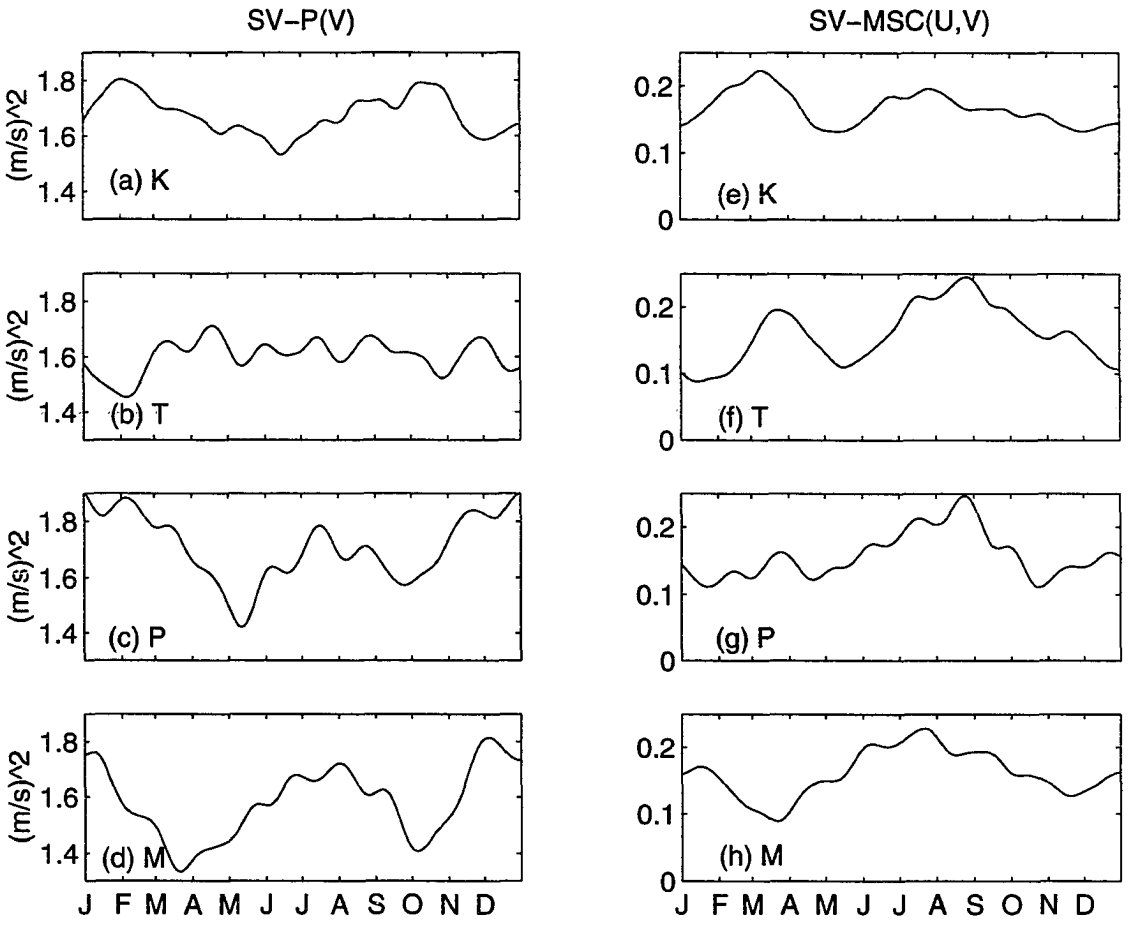

Figure 6 Seasonally varying v-wind variance $(ms^{-1})^2$ (a - d) and MSC between u- and v-winds (e - h) at Koror (a,e), Truk (b,f), Ponape (c,g), and Majuro (d,h).

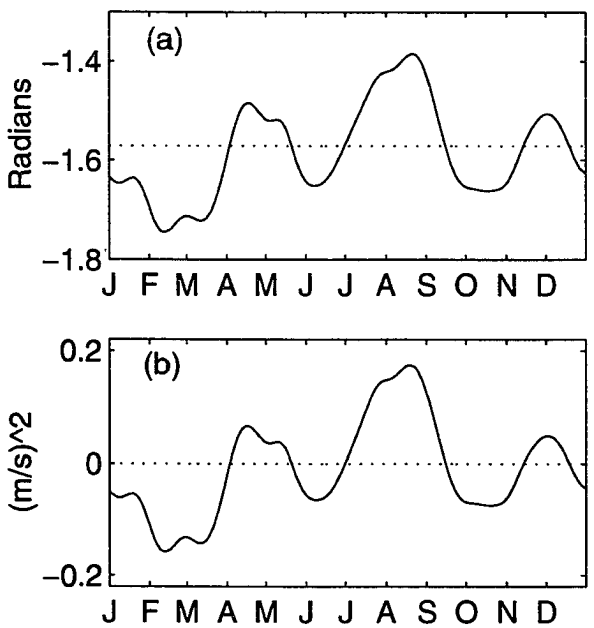

Figure 7 Seasonally varying (a) MSC phase between u- and v-wind (radians), and (b) SV covariance $(ms^{-1})^2$ at Koror.

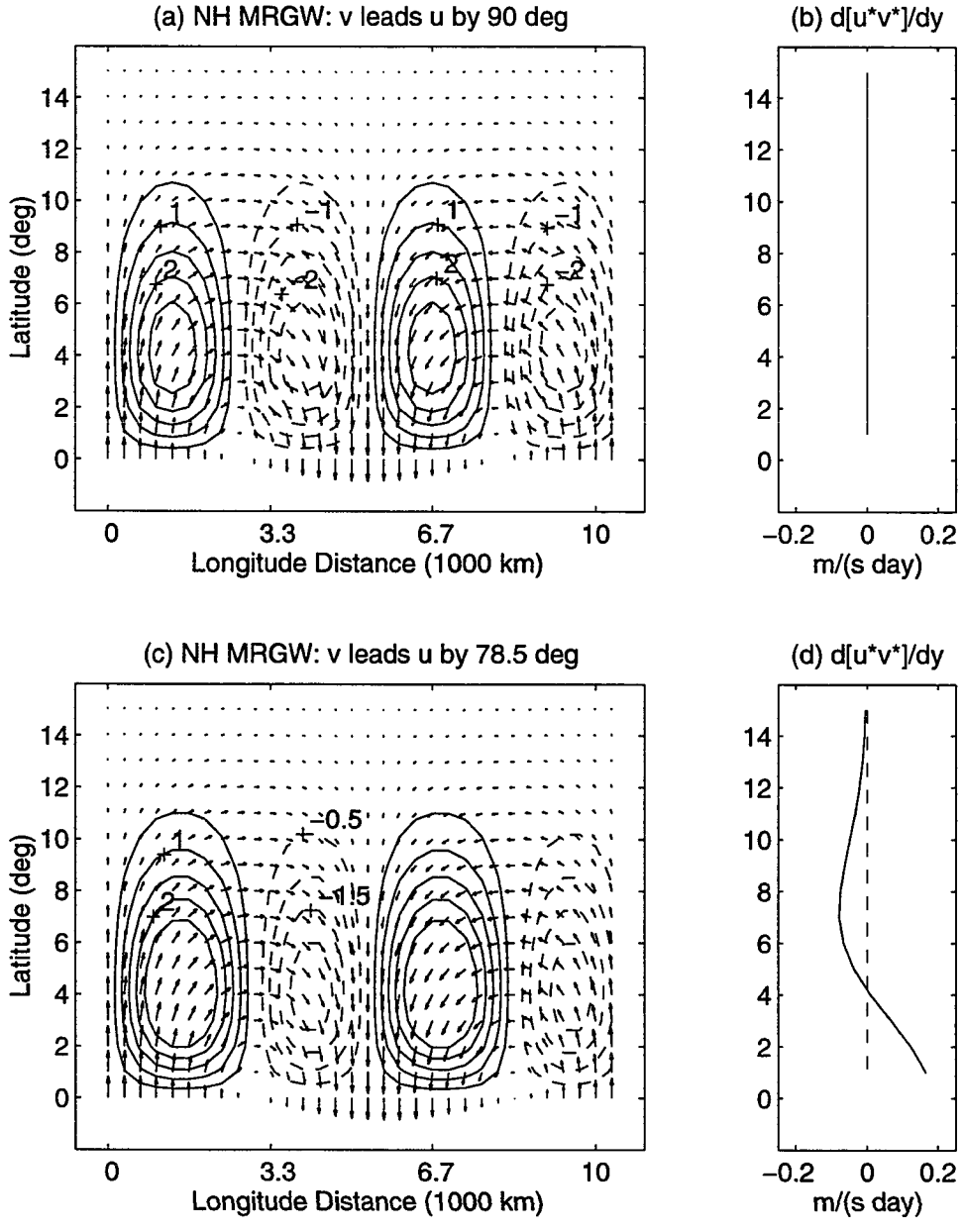

Figure 8 (a) NH half of a theoretical mixed Rossby-gravity wave (MRGW) wind field (i.e., v leads u by 90 degrees) with contours of horizontal eddy momentum flux (u^*v^*) (contour interval .5 $(ms^{-1})^2$). (b) Meridional derivative of horizontal momentum flux ($ms^{-1}day^{-1}$) corresponding to the wave in figure (a). (c) and (d) are the same as for (a) and (b) except the MRGW v-wind leads the u-wind by only 78.5 degrees.

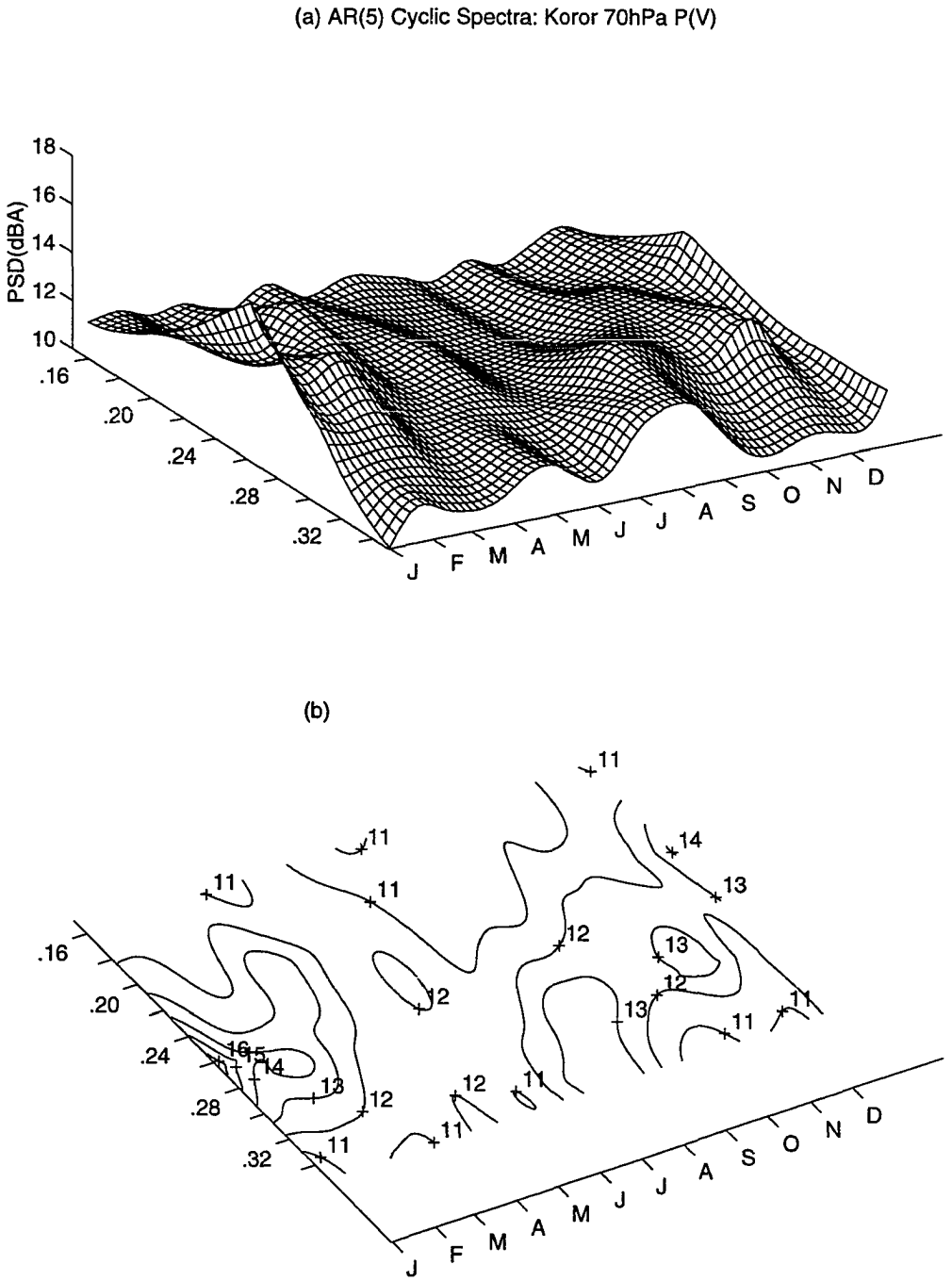

Figure 9 Autoregressive cyclic spectra (assuming a constant AR model order equal to 5) for 70-hPa v-wind at Koror (in dBA). Note that the frequency domain has been truncated to the region of interest (.16 - .34 cpd) and the contour intervals are 1 dBA in (b).

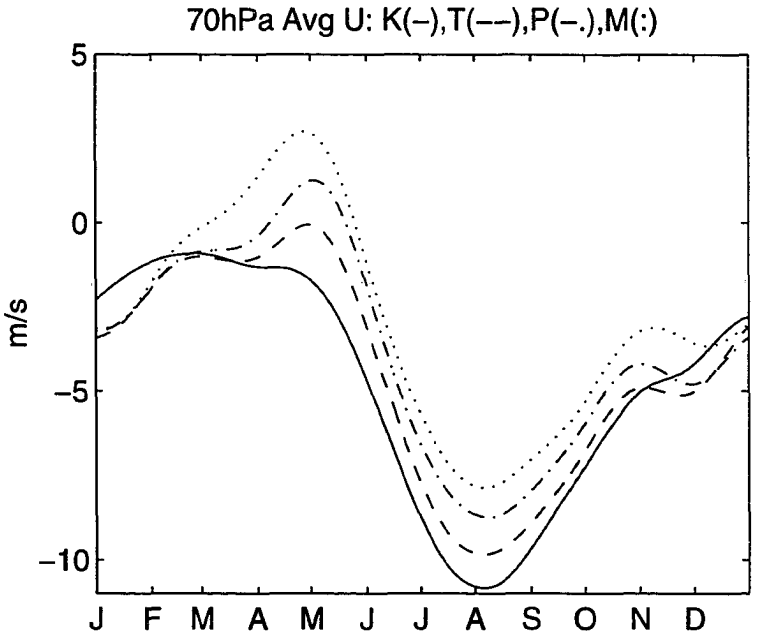

Figure 10 31-year mean zonal wind (ms^{-1}) at 70-hPa for Koror (solid line), Truk (dashed), Ponape (dash-dot), Majuro (dotted). The daily means have been smoothed with a Butterworth (order 3) lowpass filter with a half-power period of 60 days.

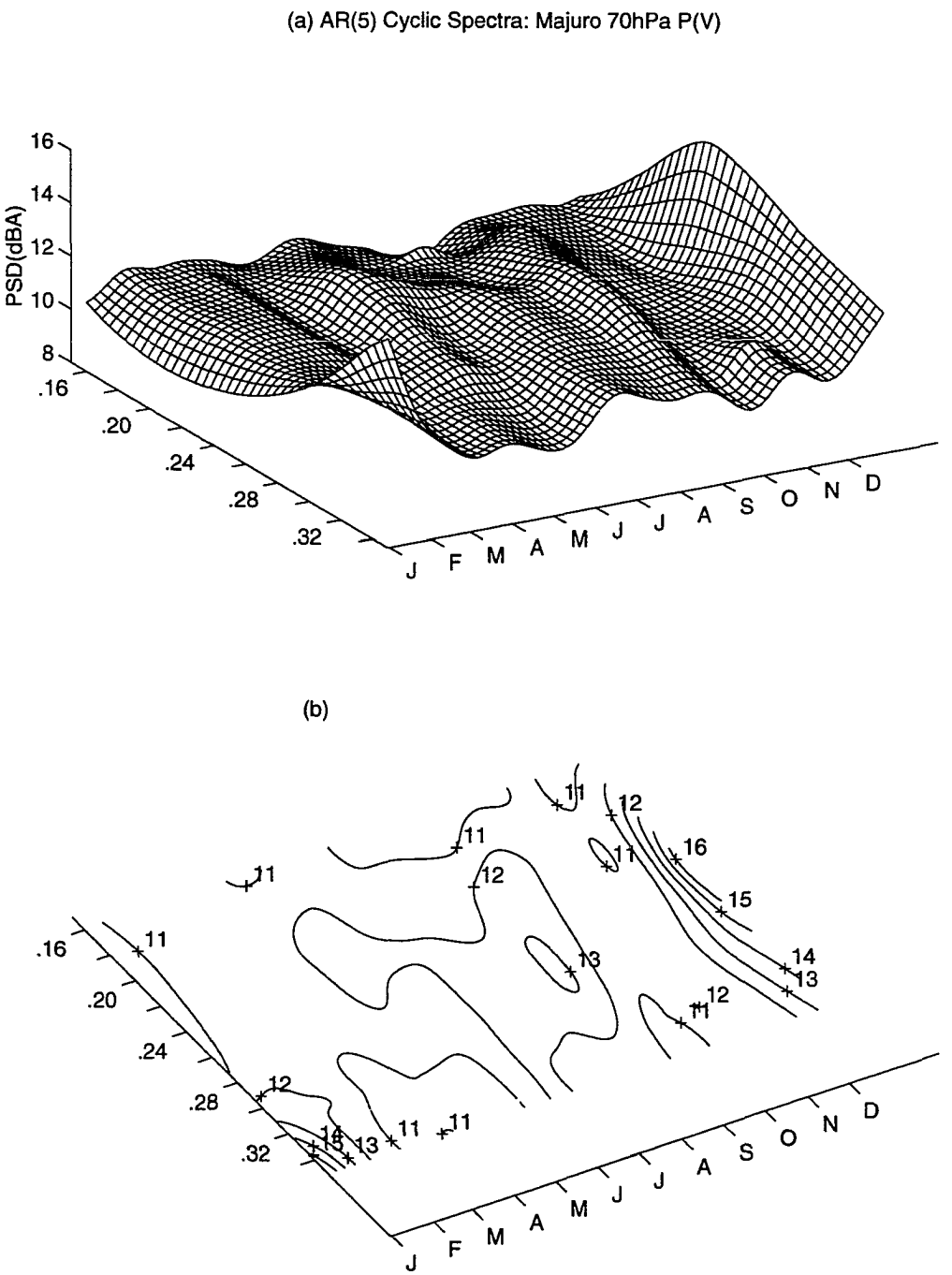

Figure 11 Autoregressive cyclic spectra (assuming a constant AR model order equal to 5) for 70-hPa v-wind at Majuro (in dBA). Note that the frequency domain has been truncated to the region of interest (.16 - .34 cpd) and the contour intervals are 1 dBA in (b).

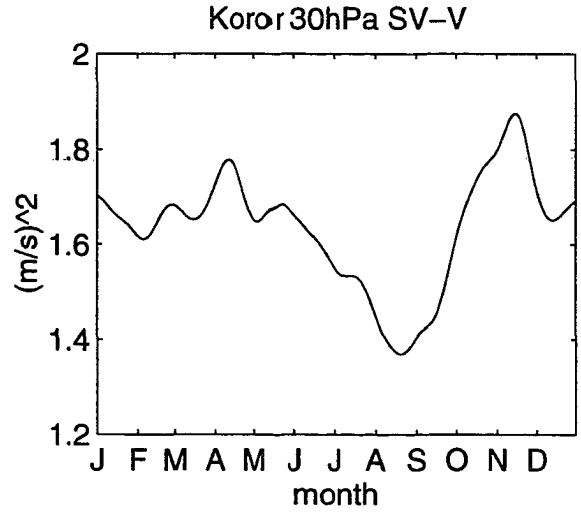

Figure 12 Seasonally varying v-wind variance (ms^{-1}) for Koror at 30-hPa.

A SPATIALLY DESCRIPTIVE, TEMPORALLY DYNAMIC STATISTICAL MODEL WITH APPLICATIONS TO ATMOSPHERIC PROCESSES

A paper - parts of which will be submitted to the
Journal of the American Statistical Association and
Monthly Weather Review

Christopher K. Wikle and Noel Cressie

Abstract

Most climatological processes involve variability over both space and time. The extension of traditional geostatistical methods, such as kriging, to the spatio-temporal domain is one possible approach to characterize this variability. Due to the difficulty in modeling space, time, and spatio-temporal interactions, this approach is limited. In the atmospheric sciences, traditional methods for examining spatio-temporal processes have focused on Empirical Orthogonal Functions (EOF), Canonical Correlation Analysis (CCA), and Principal Oscillation Patterns (POP). Although these techniques are visually powerful, they were designed with summarization rather than prediction in mind. Our predictive model is temporally dynamic in that it exploits the unidirectional flow of time in an autoregressive framework. In addition, the model is spatially descriptive in the sense that although spatial correlation is modeled by a spatially colored noise process, no causative interpretation is associated with this noise. With the inclusion of a measurement equation, this formulation naturally leads to the development

of a spatio-temporal Kalman filter. We use this Kalman filter to predict at future times and at locations for which we do not have data. The method is demonstrated with a simulated spatio-temporal data set, and is shown to perform better than applying simple kriging independently to the spatial field at each time. Finally, our approach is used to predict monthly precipitation throughout the data-sparse South China Sea region.

1 Introduction

Virtually all physical processes involve variability over space and time. For example, in climatology we are typically interested in the time evolution of certain atmospheric parameters (e.g., wind, temperature, precipitation) over specified spatial domains. In the atmospheric sciences, many methods have been developed to examine such spatio-temporal variability: Empirical Orthogonal Functions or EOFs (e.g., Lorenz 1956; Preisendorfer 1988), spatio-temporal Canonical Correlation Analysis or CCA (e.g., Glahn 1968; Bretherton et al. 1992), and Principal Oscillation Patterns or POPs (e.g., Hasselmann 1988; von Storch et al. 1988,1994). Although these techniques are visually very powerful, they were developed more as a tool with which to summarize the huge spatio-temporal data sets typically found in atmospheric science, rather than as a methodology for prediction in either space or time.

One approach to modeling spatio-temporal variability is to consider the data as separate time series, which are correlated in space (i.e., a multivariate time series model). This approach is described in Bennet (1979, ch.6) and recently has been implemented in Rouhani and Wackernagel (1990) and Oehlert (1993).

Another approach to modeling spatio-temporal variability is through the geostatistical paradigm. For instance, traditional geostatistical methods, such as kriging, can be extended to the spatio-temporal domain. However, as outlined by Rouhani and Meyers (1990), there are major differences between temporal and spatial processes (e.g., temporal data are ordered while spatial data are not). Furthermore, geostatistical spatio-temporal modeling is complicated by having to specify not only space and time components, but also spatio-temporal interaction

components of variation. The spatio-temporal variability is often complicated since there can be very different spatial behavior at different points in time, as well as different temporal variability at different locations in space.

The primary geostatistical approach has been to treat time and space as separable, so that if the time component is removed the data can be viewed as repeated measurements at each spatial location (e.g., Bilonick 1983; Eynon and Switzer 1983; Stein 1986; Loader and Switzer 1992; Sampson and Guttorp 1992; Mardia and Goodall 1993; Host et al. 1995; Haas 1995). Traditionally, spatial processes have been assumed to be covariance stationary. However, in recent years that assumption has been shown to be unrealistic for atmospheric data extending over large spatial domains. Nonstationary spatial covariances have been considered through EOFs (Obled and Creutin 1986; Shriver and O'Brien 1995; Reynolds et al. 1996), moving windows (Haas 1990a,b; 1995), kernel smoothers (Oehlert 1992), empirical Bayes shrinkage (Loader and Switzer 1992), and multidimensional scaling (Sampson and Guttorp 1992). For an excellent review of nonstationary covariance modeling, see Guttorp and Sampson (1994). Recently, Zucchini and Guttorp (1991) and Hughes and Guttorp (1994a,b) have considered using hidden Markov models with unobserved weather states to model spatio-temporal atmospheric processes.

Fundamentally, it is clear that without the spatial component, there is a large class of time series that could be used to model the temporal component (e.g., autoregressive error processes). These are *dynamic* in the sense that they exploit the unidirectional flow of time. Furthermore, without the temporal component, geostatistical methods could be used to model the spatial component (e.g., intrinsically stationary error process). These are *descriptive* in the sense that although they model spatial correlation, there is no causative interpretation associated with them. It would seem then that if both temporal and spatial components are present, it would be natural to combine both approaches. In other words, we shall propose a statistical model that is temporally dynamic and spatially descriptive. In a spatio-temporal framework where the models are separable, this has been shown to be a useful approach (Haslett

and Raftery 1989; Handcock and Wallis 1994).

In the atmospheric sciences, the idea of using a Kalman filter for spatio-temporal modeling has been discussed in the context of Numerical Weather Prediction (NWP) data initialization since the early 1980s (e.g., Ghil et al. 1981). The fact that this work has been largely ignored in the spatial statistical literature is illustrative of the lack of cross-referencing between the geostatistical "kriging" literature (Matheron 1963) and the atmospheric science "optimal interpolation" (Gandin 1963) literature noted in Haslett's (1989) excellent review. [For additional overviews of optimal interpolation in the atmospheric sciences and a discussion of the role of the Kalman filter, see Thiebaux and Pedder (1989), Daley (1991), and Ghil and Malanotte-Rizzoli (1991).] In the NWP Kalman filter, the state process is assumed to evolve according to a physical, albeit very simplified, multivariate model of the atmosphere. The approach has not been implemented operationally due to the tremendous computational costs associated with matrix operations involving, on the order of, 10^6 variables. However, Kalman filters with simplified dynamical models have been demonstrated to work quite well (e.g., Dee et al. 1985; Cohn and Parrish 1991; Dee 1991). Current research in this area is focused on using more physically realistic non-linear Kalman filters such as the "extended" Kalman filter (e.g., Miller et al. 1994; Daley 1995) and finding appropriate parameterizations of the covariance matrices associated with the error between observations and a "first guess" derived from a numerical dynamical atmospheric forecast model (e.g., Dee 1995). Although these Kalman filter models have great potential for optimally preparing atmospheric data streams to be used by physical numerical prediction models, the physically-based state matrices prevent these models from being used for spatio-temporal processes for which explicit physical models are not well understood (e.g.,precipitation). Thus it would be useful to develop a Kalman filter based on statistically derived state matrices. Such a model could then be applied to a broader class of spatio-temporal processes, namely precipitation.

Commenting on Handcock and Wallis' Bayesian approach, Cressie (1994) suggests that a Kalman filter incorporating space and time would be a powerful way to apply the Bayesian

paradigm to spatio-temporal problems. In an invited paper to the XVIIth International Biometric Society, Goodall and Mardia (1994) suggest an approach to Kalman filtering in the spatio-temporal setting, although they do not formulate their model in a continuous spatial domain. Zhang (1995) implements a spatio-temporal Kalman filter using general partial differential equation based covariance structures, with statistically-derived parameters. Huang and Cressie (1996) develop a rather specific temporally dynamic and spatially descriptive model and show how to use the Kalman filter algorithm to obtain snow-water equivalent predictions at locations where no observations are taken. They demonstrate that this spatio-temporal Kalman filter performs better than the purely spatial model currently employed by the National Weather Service for such predictions. Note, however, that for the model presented by Huang and Cressie (1996), prediction at some location $\mathbf{s}$ and time t is influenced directly by past values only at location $\mathbf{s}$. In reality, spatio-temporal processes are likely to be more complicated and will also show dependence on past values at locations *near* $\mathbf{s}$.

In this investigation, we extend the model of Huang and Cressie (1996) to include dynamical contributions from all locations in the spatial domain of interest, resulting in a spatio-temporal model that is temporally dynamic and spatially descriptive. Our approach is to express the dynamics through Markov time dependence and to describe the error through nonstationary, anisotropic, spatially colored noise. This model is formulated in a state-space representation, leading to updatable Kalman filter spatio-temporal prediction algorithms. Section 2 describes the model and the estimation of model parameters is outlined in Section 3. Section 4 describes the selection of basis functions needed for the model implementation. Then, a simulation with which we test the model and compare to other methods is included in Section 5, followed by the results of applying the model to monthly precipitation data observed around the South China Sea in Section 6. Finally, a conclusion is presented in Section 7.

2 Statistical Model

Assume we are given an observable and spatially continuous spatial process $Z(\mathbf{s};t)$, where $\mathbf{s} \in D$, with D some spatial domain in d-dimensional Euclidean space R^d, and a discrete index of times $t \in \{1,2,\ldots\}$. We suppose that the observable process has a component of measurement error expressed through the measurement equation

$$Z(\mathbf{s};t) = Y(\mathbf{s};t) + \epsilon(\mathbf{s};t), \tag{1}$$

where $Y(\mathbf{s};t)$ can be thought of as a "smoother" process than $Z(\mathbf{s};t)$. Our goal is to predict the process $Y(\cdot\,;\cdot)$. Now,we assume that $Y(\mathbf{s};t)$ can be written

$$Y(\mathbf{s};t) = Y_K(\mathbf{s};t) + \nu(\mathbf{s};t), \tag{2}$$

where $\nu(\mathbf{s};t)$ is a component of variance representing small-scale spatial variation or the convolution of spatial processes that do not have a coherent temporally dynamic structure. The component $Y_K(\mathbf{s};t)$ is assumed to evolve according to the state equation

$$Y_K(\mathbf{s};t) = \int_D w_{\mathbf{s}}(\mathbf{u}) Y_K(\mathbf{u};t-1)\,d\mathbf{u} + \eta(\mathbf{s};t), \tag{3}$$

where $\eta(\mathbf{s};t)$ is a spatially colored noise process (i.e., the "spatially descriptive" component) and $w_{\mathbf{s}}(\mathbf{u})$ is a function representing the interaction between the state process Y_K at location $\mathbf{u}$ and time $(t-1)$ and Y_K at location $\mathbf{s}$ and time t (i.e., the "temporally dynamic" component). For stationarity over time, we further require that this interaction function satisfies

$$\int_D w_{\mathbf{s}}(\mathbf{u})\,d\mathbf{u} = \alpha, \tag{4}$$

where $|\alpha| < 1$ is an unknown parameter. Additionally, we assume

$$E[\epsilon(\mathbf{s};t)Y(\mathbf{r};\tau)] = 0 \;\; \forall\, \mathbf{s},\mathbf{r},t,\tau \tag{5}$$

$$E[\epsilon(\mathbf{s};t)\epsilon(\mathbf{r};\tau)] = 0 \;\; \forall\, \mathbf{s},\mathbf{r},t \neq \tau \tag{6}$$

$$E[\nu(\mathbf{s};t)\nu(\mathbf{r};\tau)] = 0 \;\; \forall\, \mathbf{s},\mathbf{r},t \neq \tau \tag{7}$$

$$E[\eta(\mathbf{s};t)\eta(\mathbf{r};\tau)] = 0 \;\; \forall\, \mathbf{s},\mathbf{r},t \neq \tau \tag{8}$$

$$E[\epsilon(\mathbf{s};t)\eta(\mathbf{r};\tau)] = 0 \ \forall \mathbf{r},\mathbf{s},t,\tau \quad (9)$$

$$E[\epsilon(\mathbf{s};t)\nu(\mathbf{r};\tau)] = 0 \ \forall \mathbf{r},\mathbf{s},t,\tau \quad (10)$$

$$E[\nu(\mathbf{s};t)\eta(\mathbf{r};\tau)] = 0 \ \forall \mathbf{r},\mathbf{s},t,\tau \quad (11)$$

$$E[\nu(\mathbf{s};t)Y_K(\mathbf{r};t)] = 0 \ \forall \mathbf{r},\mathbf{s},t \quad (12)$$

$$E[\eta(\mathbf{s};t)Y_K(\mathbf{r};t-1)] = 0 \ \forall \mathbf{r},\mathbf{s},t. \quad (13)$$

We have incomplete (in space and time) observations on the Z process, from which we would like to predict the unobserved process $Y(\mathbf{s}_0;t_0)$, where $\mathbf{s}_0$ and t_0 may or may not represent spatio-temporal coordinates at which data are available. The state model (3) is an extension of the model given by Huang and Cressie (1996) who effectively assume that $\nu(\cdot\,;t) \equiv 0$ and that $w_{\mathbf{s}}(\mathbf{u})$ is an unknown parameter times the Dirac delta function, thereby only considering contributions to $Y_K(\mathbf{s};t)$ from previous values of the process at the *same* location $\mathbf{s}$. The strength of the state model (3) is that it features the dynamic aspect through the continuous autoregressive term but builds in spatio-temporal interaction through the error process $\eta(\cdot\,;t)$, which is, at any point in time, a spatially correlated (e.g., intrinsically stationary) process.

Now, we define the state process Y_K according to

$$Y_K(\mathbf{s};t) \equiv \sum_{k=1}^{K} \phi_k(\mathbf{s})a_k(t), \quad (14)$$

for $\mathbf{s} \in D, t \in \{1,2,\ldots\}$, where $a_k(t), k = 1,\ldots,K$ are zero-mean random variables, and $\{\phi_i(\mathbf{u}) : i = 1,\ldots,\infty; \mathbf{u} \in D\}$ is some chosen basis set that is complete (i.e., for any piecewise continuous function $f(\mathbf{u}), \mathbf{u} \in D$, the minimum squared error of $f(\mathbf{u}) - \sum_{i=1}^{m} c_i\phi_i(\mathbf{u})$ on the appropriate vector space goes to zero as $m \to \infty$) and orthonormal,

$$\int_D \phi_i(\mathbf{u})\phi_j(\mathbf{u})d\mathbf{u} = \delta_{i,j}, \quad (15)$$

where $\delta_{i,j} = 0$ for $i \neq j$ and $\delta_{i,j} = 1$ for $i = j$. Choice of the integer $K \geq 1$ will be discussed in Section 3. Furthermore, we take advantage of the completeness of the ϕ's and expand the interaction function as

$$w_{\mathbf{s}}(\mathbf{u}) = \sum_{l=1}^{\infty} b_l(\mathbf{s})\phi_l(\mathbf{u}), \quad (16)$$

for $\mathbf{s}, \mathbf{u} \in D$ and where $b_l(\mathbf{s}), l = 1, \ldots, \infty$ are unknown, but *non-stochastic*, parameters. Notice that (15) implies

$$b_l(\mathbf{s}) = \int_D w_{\mathbf{s}}(\mathbf{u})\phi_l(\mathbf{u})d\mathbf{u}, \tag{17}$$

for $\mathbf{s}, \mathbf{u} \in D$.

Substituting (16) and (14) into (3) and making use of the orthonormality property (15) we obtain

$$\begin{aligned} Y_K(\mathbf{s};t) &= \sum_{k=1}^{K}\sum_{l=1}^{\infty} b_l(\mathbf{s})a_k(t-1)\int_D \phi_l(\mathbf{u})\phi_k(\mathbf{u})d\mathbf{u} + \eta(\mathbf{s};t) && (18)\\ &= \sum_{k=1}^{K} b_k(\mathbf{s})a_k(t-1) + \eta(\mathbf{s};t) && (19)\\ &= \mathbf{b}(\mathbf{s})'\mathbf{a}(t-1) + \eta(\mathbf{s};t), && (20) \end{aligned}$$

where

$$\begin{aligned} \mathbf{b}(\mathbf{s}) &\equiv (b_1(\mathbf{s}), b_2(\mathbf{s}), \ldots, b_K(\mathbf{s}))', && (21)\\ \mathbf{a}(t-1) &\equiv (a_1(t-1), a_2(t-1), \ldots, a_K(t-1))'. && (22) \end{aligned}$$

Noting that (14) can be written

$$Y_K(\mathbf{s};t) = \boldsymbol{\phi}(\boldsymbol{s})'\mathbf{a}(t), \tag{23}$$

where

$$\boldsymbol{\phi}(\boldsymbol{s}) \equiv (\phi_1(\mathbf{s}), \phi_2(\mathbf{s}), \ldots, \phi_K(\mathbf{s}))', \tag{24}$$

we substitute (23) into (20) to obtain

$$\boldsymbol{\phi}(s)'\mathbf{a}(t) = \mathbf{b}(\mathbf{s})'\mathbf{a}(t-1) + \eta(\mathbf{s};t). \tag{25}$$

Now, assume that we have data at locations $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ for t in some non-empty subset of $\{1, 2, \ldots, T\}$. We can then write (25) as a system of n equations, one for each $\mathbf{s}_i, i = 1, \ldots, n$. In matrix form, this linear system is written as

$$\boldsymbol{\Phi}\mathbf{a}(t) = \mathbf{B}\mathbf{a}(t-1) + \boldsymbol{\eta}(t), \tag{26}$$

where

$$\boldsymbol{\eta}(t) \equiv (\eta(\mathbf{s}_1;t), \eta(\mathbf{s}_2;t), \ldots, \eta(\mathbf{s}_n;t))', \tag{27}$$

and where we define the $n \times K$ matrices

$$\boldsymbol{\Phi} \equiv (\boldsymbol{\phi}(s_1), \boldsymbol{\phi}(s_2), \ldots, \boldsymbol{\phi}(s_n))', \tag{28}$$

$$\mathbf{B} \equiv (\mathbf{b}(\mathbf{s}_1), \mathbf{b}(\mathbf{s}_2), \ldots, \mathbf{b}(\mathbf{s}_n))'. \tag{29}$$

Assuming $n \geq K$ and $[\boldsymbol{\Phi}'\boldsymbol{\Phi}]^{-1}$ is non-singular, we can write (25) as

$$\mathbf{a}(t) = [\boldsymbol{\Phi}'\boldsymbol{\Phi}]^{-1}\boldsymbol{\Phi}'\mathbf{B}\mathbf{a}(t-1) + [\boldsymbol{\Phi}'\boldsymbol{\Phi}]^{-1}\boldsymbol{\Phi}'\boldsymbol{\eta}(t) \tag{30}$$

$$= \mathbf{J}\mathbf{B}\mathbf{a}(t-1) + \boldsymbol{J}\boldsymbol{\eta}(t) \tag{31}$$

$$= \mathbf{H}\mathbf{a}(t-1) + \boldsymbol{J}\boldsymbol{\eta}(t), \tag{32}$$

where we have defined the $K \times n$ matrix

$$\mathbf{J} \equiv [\boldsymbol{\Phi}'\boldsymbol{\Phi}]^{-1}\boldsymbol{\Phi}' \tag{33}$$

and the $K \times K$ matrix

$$\mathbf{H} \equiv \mathbf{J}\mathbf{B}. \tag{34}$$

We can then rewrite the measurement equation (1) as

$$Z(\mathbf{s};t) = \boldsymbol{\phi}(\boldsymbol{s})'\mathbf{a}(t) + \nu(\mathbf{s};t) + \epsilon(\mathbf{s};t). \tag{35}$$

We note that as a consequence of the original model assumptions, we have

$$\mathrm{E}[a_i(t)\nu(\mathbf{s},\tau)] = 0, \quad \text{for all } i = 1, \ldots, K; \mathbf{s}, t, \tau \tag{36}$$

$$\mathrm{E}[a_i(t)\epsilon(\mathbf{s},\tau)] = 0, \quad \text{for all } i = 1, \ldots, K; \mathbf{s}, t, \tau. \tag{37}$$

2.1 Kalman Filter Representation

Recall from (23) that $Y_K(\mathbf{s};t) = \boldsymbol{\phi}(\boldsymbol{s})'\mathbf{a}(t)$ and hence, if an optimal (minimum mean-squared prediction error) predictor of $\mathbf{a}(t)$ is found, the optimal predictor of $Y_K(\mathbf{s};t)$ is immediately available after premultiplying by $\boldsymbol{\phi}(\mathbf{s})'$. It is then a simple matter to obtain the optimal

predictor for $Y(\mathbf{s};t)$, as we shall see below. The optimal predictor of $\mathbf{a}(t)$ given observations up to and including t is

$$\hat{\mathbf{a}}(t \mid t) \equiv \mathrm{E}[\mathbf{a}(t) \mid \mathbf{Z}(t), \ldots, \mathbf{Z}(1)], \tag{38}$$

for $t \geq 1$, with mean-squared prediction error,

$$\mathbf{P}(t \mid t) \equiv \mathrm{E}\{[\mathbf{a}(t) - \hat{\mathbf{a}}(t \mid t)][\mathbf{a}(t) - \hat{\mathbf{a}}(t \mid t)]'\}. \tag{39}$$

These quantities can be calculated recursively by means of a Kalman filter as derived in Appendix A. We obtain the following recursion equations for $t \geq 1$:

$$\hat{\mathbf{a}}(t \mid t) = \hat{\mathbf{a}}(t \mid t-1) + \mathbf{K}(t)[\mathbf{Z}(t) - \mathbf{\Phi}\hat{\mathbf{a}}(t \mid t-1)] \tag{40}$$

$$\mathbf{P}(t \mid t) = \mathbf{P}(t \mid t-1) - \mathbf{K}(t)\mathbf{\Phi}\mathbf{P}(t \mid t-1), \tag{41}$$

where $\mathbf{Z}(t) = (Z(\mathbf{s}_1;t), \ldots, Z(\mathbf{s}_n;t))'$ and the Kalman gain $\mathbf{K}(t)$ is given by

$$\mathbf{K}(t) = \mathbf{P}(t \mid t-1)\mathbf{\Phi}'[\mathbf{R} + \mathbf{V} + \mathbf{\Phi}\mathbf{P}(t \mid t-1)\mathbf{\Phi}']^{-1}, \tag{42}$$

with one-step ahead predictions given by

$$\hat{\mathbf{a}}(t \mid t-1) \equiv \mathrm{E}[\mathbf{a}(t) \mid \mathbf{Z}(t-1), \ldots, \mathbf{Z}(1)] \tag{43}$$

$$= \mathbf{H}\hat{\mathbf{a}}(t-1 \mid t-1) \tag{44}$$

$$\mathbf{P}(t \mid t-1) \equiv \mathrm{var}[\mathbf{a}(t) \mid \mathbf{Z}(t-1), \ldots, \mathbf{Z}(1)] \tag{45}$$

$$= \mathbf{H}\mathbf{P}(t-1 \mid t-1)\mathbf{H}' + \mathbf{J}\mathbf{Q}\mathbf{J}', \tag{46}$$

where

$$\mathbf{R} \equiv \mathrm{var}[\epsilon(t)] \tag{47}$$

$$\mathbf{V} \equiv \mathrm{var}[\nu(t)] \tag{48}$$

$$\mathbf{Q} \equiv \mathrm{var}[\eta(t)], \tag{49}$$

and where

$$\epsilon(t) \equiv (\epsilon(\mathbf{s}_1;t), \ldots, \epsilon(\mathbf{s}_n;t))'$$

$$\nu(t) \equiv (\nu(\mathbf{s}_1;t), \ldots, \nu(\mathbf{s}_n;t))'.$$

Now, consider prediction of the process $Y(\mathbf{s};t)$ given the Kalman filter predictor $\hat{\mathbf{a}}(t \mid t)$. Assuming multivariate normality as in Appendix A, the optimal predictor is then derived as follows:

$$\begin{aligned}
\hat{Y}(\mathbf{s};t \mid t) &= \mathrm{E}[Y(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] && (50)\\
&= \mathrm{E}[\boldsymbol{\phi}(\mathbf{s})'\mathbf{a}(t) + \nu(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] && (51)\\
&= \boldsymbol{\phi}(\mathbf{s})'\mathrm{E}[\mathbf{a}(t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] + \mathrm{E}[\nu(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] && (52)\\
&= \boldsymbol{\phi}(\mathbf{s})'\hat{\mathbf{a}}(t \mid t) + \mathbf{c}_\nu(\mathbf{s})'[\mathbf{C}_0^Z]^{-1}\mathbf{Z}(t), && (53)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{Z}^*(t-1) &\equiv [\mathbf{Z}(t-1), \ldots, \mathbf{Z}(1)], && (54)\\
\mathbf{C}_0^Z &\equiv \mathrm{cov}[\mathbf{Z}(t), \mathbf{Z}(t)], && (55)\\
\mathbf{c}_\nu(\mathbf{s}) &\equiv \mathrm{E}[\nu(\mathbf{s};t)\boldsymbol{\nu}(t)] && (56)\\
&= (c_\nu(\mathbf{s}, \mathbf{s}_1), \ldots c_\nu(\mathbf{s}, \mathbf{s}_n))', && (57)
\end{aligned}$$

and where

$$c_\nu(\mathbf{s}, \mathbf{r}) \equiv \mathrm{E}[\nu(\mathbf{s};t)\nu(\mathbf{r};t)]. \quad (58)$$

We note that the second term in (53) is a type of simple kriging (e.g., Cressie 1993, p.110) or optimal smoothing applied to the residual spatial noise term $\nu(\mathbf{s};t)$. Thus, as the truncation integer K approaches one, the optimal predictor of $Y(\mathbf{s};t)$ begins to look more and more like the simple kriging predictor in the presence of measurement error. In practice, the truncation K is generally large, implying that the contribution to (53) from the ν process, and hence the associated simple kriging predictor, is relatively small.

The prediction error variance for $Y(\mathbf{s};t)$, assuming multivariate normality, is given by

$$\begin{aligned}
\mathrm{var}[Y(\mathbf{s};t) - \hat{Y}(\mathbf{s};t \mid t)] &= \mathrm{var}[Y(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] && (59)\\
&= \mathrm{var}[\boldsymbol{\phi}(\mathbf{s})'\mathbf{a}(t) + \nu(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] && (60)\\
&= \mathrm{var}[\boldsymbol{\phi}(\mathbf{s})'\mathbf{a}(t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] +
\end{aligned}$$

$$\begin{aligned}
& \operatorname{var}[\nu(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] + \\
& 2\operatorname{cov}[\boldsymbol{\phi}(\mathbf{s})'\mathbf{a}(t), \nu(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] \qquad (61) \\
= \; & \boldsymbol{\phi}(\mathbf{s})'\mathbf{P}(t \mid t)\boldsymbol{\phi}(\mathbf{s}) + c_\nu(\mathbf{s},\mathbf{s}) - \mathbf{c}_\nu(\mathbf{s})'[\mathbf{C}_0^Z]^{-1}\mathbf{c}_\nu(\mathbf{s}) \\
& +2\operatorname{cov}[\boldsymbol{\phi}(\mathbf{s})'\mathbf{a}(t), \nu(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] \qquad (62) \\
= \; & \boldsymbol{\phi}(\mathbf{s})'\mathbf{P}(t \mid t)\boldsymbol{\phi}(\mathbf{s}) + c_\nu(\mathbf{s},\mathbf{s}) - \mathbf{c}_\nu(\mathbf{s})'[\mathbf{C}_0^Z]^{-1}\mathbf{c}_\nu(\mathbf{s}) \\
& -2\,\boldsymbol{\phi}(\mathbf{s})'\operatorname{cov}[\hat{\mathbf{a}}(t \mid t), \mathbf{Z}(t)][\mathbf{C}_0^Z]^{-1}\mathbf{c}_\nu(\mathbf{s}), \qquad (63)
\end{aligned}$$

where, assuming Gaussian processes, we have used the relationship

$$\begin{aligned}
\operatorname{cov}[\boldsymbol{\phi}(\mathbf{s})'\mathbf{a}(t), \nu(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] &= -\operatorname{cov}\{\mathrm{E}[\boldsymbol{\phi}(\mathbf{s})'\mathbf{a}(t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)], \\
& \qquad \mathrm{E}[\nu(\mathbf{s};t) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)]\} \qquad (64) \\
&= -\operatorname{cov}\{\boldsymbol{\phi}(\mathbf{s})'\hat{\mathbf{a}}(t \mid t), \mathbf{c}_\nu(\mathbf{s})'[\mathbf{C}_0^Z]^{-1}\mathbf{Z}(t)\} \qquad (65) \\
&= -\boldsymbol{\phi}(\mathbf{s})'\operatorname{cov}[\hat{\mathbf{a}}(t \mid t), \mathbf{Z}(t)][\mathbf{C}_0^Z]^{-1}\mathbf{c}_\nu(\mathbf{s}). \qquad (66)
\end{aligned}$$

Note that the first term in (63) corresponds to the prediction error variance from the Y_K process, while second and third terms in (63) correspond to the simple kriging variance (e.g., Cressie 1993, p.110) of the ν process. The last term in (63) is then a correction due to the covariance between the Kalman filter prediction of the Y_K process (through $a(\cdot)$) and the simple kriging predictor. It is clear that as the truncation integer K goes to one, the prediction error variance looks more and more like the simple kriging variance.

Similarly, the one-step ahead optimal predictor and prediction error variance for $Y(\mathbf{s};t+1)$ are, respectively:

$$\begin{aligned}
\hat{Y}(\mathbf{s};t+1 \mid t) &= \mathrm{E}[Y(\mathbf{s};t+1) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] \qquad (67) \\
&= \boldsymbol{\phi}(\mathbf{s})'\hat{\mathbf{a}}(t+1 \mid t) \qquad (68) \\
\operatorname{var}[Y(\mathbf{s};t+1) - \hat{Y}(\mathbf{s};t \mid t)] &= \operatorname{var}[Y(\mathbf{s};t+1) \mid \mathbf{Z}(t), \mathbf{Z}^*(t-1)] \qquad (69) \\
&= \boldsymbol{\phi}(\mathbf{s})'\mathbf{P}(t+1 \mid t)\boldsymbol{\phi}(\mathbf{s}) + c_\nu(\mathbf{s},\mathbf{s}), \qquad (70)
\end{aligned}$$

where the one-step ahead predictors $\hat{\mathbf{a}}(t+1 \mid t)$ and $\mathbf{P}(t+1 \mid t)$ are given by (44) and (46), respectively. Note that the independence between the ν process at $(t+1)$ and the $\mathbf{Z}$'s at previous

times accounts for the simplification of the one-step ahead predictors relative to the predictors for $Y(\mathbf{s};t)$. That is, the kriging terms do not contribute to the one-step ahead prediction since the ν process is assumed not to evolve coherently in time.

3 Estimation of Model Parameters

The Kalman filter presented in Section 2.1 gives optimal predictors only if we know the true error covariances $\mathbf{R}, \mathbf{V}$, and $\mathbf{Q}$, as well as the state matrix $\mathbf{H}$ (i.e., Φ and $\mathbf{B}$). In our case, we can *choose* the ϕ's, as long as they are complete and orthonormal, but we must *estimate* the $\mathbf{R}, \mathbf{V}, \mathbf{Q}$ and $\mathbf{B}$ matrices. Although we lose optimality by not knowing these matrices, our model is more broadly applicable to different physical processes if we let the data determine the structure of these matrices. This corresponds to viewing the Kalman filter as an empirical Bayesian technique. Although maximum likelihood (ML) estimators of model parameters are more efficient, the high-dimensional nature of spatio-temporal problems makes for poorly behaved likelihood surfaces and iterative ML solutions that are difficult to implement. We thus focus on the simpler method of moments (MOM) estimators.

3.1 Estimation of Model Covariances

This section describes the estimation of the $\mathbf{R}, \mathbf{V}$, and $\mathbf{Q}$ error covariance matrices, as defined in (47)-(49), respectively.

3.1.1 Estimation of R

In (6) we have assumed that measurement error is spatial white noise so that

$$\mathrm{E}[\epsilon(\mathbf{s};t)\epsilon(\mathbf{r};t)] = \sigma_\epsilon^2, \quad \text{for } \mathbf{s} = \mathbf{r}, \tag{71}$$

and is zero when $\mathbf{s} \neq \mathbf{r}$, for all t. Then,

$$\mathbf{R} = \sigma_\epsilon^2 \mathbf{I}, \tag{72}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. We can estimate σ_ϵ^2, preferably from information about the measuring instrument, or through the behavior of an empirical variogram estimate of the data $\{\mathbf{Z}(t), \ldots, \mathbf{Z}(1)\}$ as the spatial lag approaches zero. [For a discussion on variograms, their definition, estimation, and modeling, see Cressie (1993, Section 2.4).] The empirical variogram is obtained from the MOM estimator

$$2\hat{\gamma}(\mathbf{h};t) \equiv \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i;t) - Z(\mathbf{s}_j;t))^2, \quad \mathbf{h} \in R^d, \tag{73}$$

where $N(\mathbf{h}) \equiv \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}; i, j = 1, \ldots, n\}$ and $|N(\mathbf{h})|$ is the number of distinct pairs in $N(\mathbf{h})$ (Cressie 1993, p.69). As described in Cressie(1993, p.74-76), it is also prudent to examine a robust estimator of the variogram to check for possible contamination by outliers. In the presence of measurement error and microscale spatial variability, the variogram estimate will not go to zero as $\mathbf{h}$ goes to zero, but instead approaches a value known in geostatistics as the "nugget effect". In the absence of extra information on the precision of the measuring instrument, we declare our measurement error estimate $\hat{\sigma}_\epsilon^2$ to be equal to this nugget effect. Then,

$$\hat{\mathbf{R}} = \hat{\sigma}_\epsilon^2 \mathbf{I}. \tag{74}$$

3.1.2 Estimation and Modeling of V

In addition to the estimate of $\mathbf{V}$ (48) needed for the Kalman filter, (53) and (63) show that we need an estimate of $\mathbf{c}_\nu(\mathbf{s})$, where $\mathbf{s}$ may be at a location where we do not have data. This suggests that we should *model* the covariance structure of the ν process. To do this, we first obtain the spectral decomposition of

$$\mathbf{C}_0^Y \equiv \text{cov}[\mathbf{Y}(t), \mathbf{Y}(t)], \tag{75}$$

given the truncated set of basis functions at data locations $\mathbf{\Phi}$. Define

$$\mathbf{L} \equiv [\mathbf{\Phi}'\mathbf{\Phi}]^{-1}\mathbf{\Phi}\mathbf{C}_0^Y\mathbf{\Phi}[\mathbf{\Phi}'\mathbf{\Phi}]^{-1} \tag{76}$$

$$= \mathbf{J}\mathbf{C}_0^Y\mathbf{J}'. \tag{77}$$

Transforming from the spectral domain back to physical space gives the covariance accounted for by the truncated basis set which is, by definition, the covariance matrix of the Y_K process at observation locations. That is,

$$\mathbf{C}_0^{Y_K} \equiv \mathbf{\Phi L \Phi}'. \tag{78}$$

Then, the residual covariance matrix associated with the ν process is given by

$$\mathbf{V} = \mathbf{C}_0^{Y} - \mathbf{C}_0^{Y_K}. \tag{79}$$

In practice, given an estimate of $\mathbf{C}_0^Y$, we can obtain an estimate of $\mathbf{V}$, which we denote $\hat{\mathbf{V}}$.

From $\hat{\mathbf{V}}$ we are only able to obtain estimates of the ν-process covariances at locations for which we have data. In order to get estimates at locations where we do not have data (which are needed for the optimal predictor of $Y(\mathbf{s};t)$), we model $c_\nu(\mathbf{s},\mathbf{r})$. This modeling is accomplished through some valid (i.e., positive-definite) covariance function (e.g., Thiebaux 1976; Cressie 1993, p.84-86). We assume a spatially stationary and isotropic covariance function (i.e., the covariance is only a function of the spatial "lag") $c_\nu(h;\boldsymbol{\theta})$, where h is the spatial lag, and $\boldsymbol{\theta}$ is a vector of model parameters. In order to estimate these model parameters, we must fit the model to $\hat{c}_\nu(h)$, empirical estimates of the covariance. We use a method-of-moments estimator of the form:

$$\hat{c}_\nu(h) \equiv \sum_{N(h)} \hat{\nu}(\mathbf{s}_i,\mathbf{s}_j)/|N(h)| \ , \tag{80}$$

where $N(h) \equiv \{(\mathbf{s}_i,\mathbf{s}_j) : ||\mathbf{s}_i - \mathbf{s}_j|| \in T(h)\}$, $|N(h)|$ is the number of distinct elements in $N(h)$, $T(h)$ is a tolerance region around h, and $\hat{\nu}(\mathbf{s}_i,\mathbf{s}_j)$ is the (i,j)-th element of $\hat{\mathbf{V}}$. The candidate model $c_\nu(h;\boldsymbol{\theta})$ can be fit to the estimator (80) by several techniques (see, e.g., Cressie 1993; Section 2.6). A weighted least squares approach represents a compromise between efficiency and simplicity and can be implemented by minimizing,

$$\sum_l \frac{[\hat{c}_\nu(h_l) - c_\nu(h_l;\boldsymbol{\theta})]^2 |N(h_l)|}{[c_\nu(0;\boldsymbol{\theta}) - c_\nu(h_l;\boldsymbol{\theta})]^2}, \tag{81}$$

with respect to the parameters $\boldsymbol{\theta}$, and where h_l denotes the spatial lags at which the estimator (80) was obtained. The motivation for this weighted least squares approach can be found in Cressie(1993, p.99).

In the case where $\hat{\mathbf{V}}$ is a diagonal matrix (or possibly diagonally dominant), we can assume that ν is simply white noise,

$$c_\nu(\mathbf{s},\mathbf{r}) = \sigma_\nu^2, \tag{82}$$

for $\mathbf{s} = \mathbf{r}$, and zero otherwise. In that case, we estimate σ_ν^2 according to

$$\hat{\sigma}_\nu^2 = (1/n)\sum_{i=1}^{n} \hat{\nu}(\mathbf{s}_i,\mathbf{s}_i), \tag{83}$$

where $\hat{\nu}(\mathbf{s}_i,\mathbf{s}_i)$ is the i-th diagonal element of $\hat{\mathbf{V}}$.

3.1.3 Estimation of Q

Examination of the Kalman filter equations (40)-(46) shows that we need only estimate $\mathbf{JQJ}'$, rather than $\mathbf{Q}$. From (32) we can write

$$\begin{aligned}
\mathbf{JQJ}' &= \mathrm{E}[(\mathbf{a}(t) - \mathbf{Ha}(t-1))(\mathbf{a}(t) - \mathbf{Ha}(t-1))'] && (84)\\
&= \mathrm{E}[\mathbf{a}(t)\mathbf{a}(t)'] - \mathrm{E}[\mathbf{a}(t)\mathbf{a}(t-1)']\mathbf{H}' \\
&\quad -\mathbf{H}\mathrm{E}[\mathbf{a}(t-1)\mathbf{a}(t)'] + \mathbf{H}\mathrm{E}[\mathbf{a}(t-1)\mathbf{a}(t-1)']\mathbf{H}' && (85)\\
&= \mathbf{J}[\mathbf{C}_0^Z - \mathbf{V} - \mathbf{R}]\mathbf{J}' - \mathbf{JC}_1^Z\mathbf{J}'\mathbf{H}' \\
&\quad -\mathbf{HJ}[\mathbf{C}_1^Z]'\mathbf{J}' + \mathbf{HJ}[\mathbf{C}_0^Z - \mathbf{V} - \mathbf{R}]\mathbf{J}'\mathbf{H}', && (86)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{C}_0^Z &\equiv \mathrm{cov}[\mathbf{Z}(t),\mathbf{Z}(t)] && (87)\\
\mathbf{C}_1^Z &\equiv \mathrm{cov}[\mathbf{Z}(t),\mathbf{Z}(t-1)], && (88)
\end{aligned}$$

and we have used the relationship derived from the vector form of (35) to obtain:

$$\begin{aligned}
\mathrm{E}[\mathbf{a}(t)\mathbf{a}(t)'] &= \mathbf{J}\mathrm{E}[(\mathbf{Z}(t) - \boldsymbol{\nu}(t) - \boldsymbol{\epsilon}(t))(\mathbf{Z}(t) - \boldsymbol{\nu}(t) - \boldsymbol{\epsilon}(t))']\mathbf{J}' && (89)\\
&= \mathbf{J}[\mathbf{C}_0^Z - \mathbf{V} - \mathbf{R}]\mathbf{J}' && (90)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}[\mathbf{a}(t)\mathbf{a}(t-1)'] &= \mathbf{J}\mathrm{E}[(\mathbf{Z}(t) - \boldsymbol{\nu}(t) - \boldsymbol{\epsilon}(t))(\mathbf{Z}(t-1) - \boldsymbol{\nu}(t-1) - \boldsymbol{\epsilon}(t-1))']\mathbf{J}' && (91)\\
&= \mathbf{JC}_1^Z\mathbf{J}', && (92)
\end{aligned}$$

where we have assumed temporal invariance (e.g., lag-zero and lag-one temporal covariances do not depend on t).

Thus, to obtain an estimate for $\mathbf{J}\hat{\mathbf{Q}}\mathbf{J}'$ we can substitute estimates for $\mathbf{C}_0^Z$,$\mathbf{C}_1^Z$, $\mathbf{V}, \mathbf{R}$, and $\mathbf{H}$ into (86), where care is taken to ensure positive definiteness of the matrices.

3.2 Estimation of State Matrix B

From (2) and (26), we can write

$$\mathbf{Y}(t) = \mathbf{B}\mathbf{a}(t-1) + \boldsymbol{\eta}(t) + \boldsymbol{\nu}(t), \tag{93}$$

for $t \geq 1$. Post-multiplying (93) by $\mathbf{Y}(t-1)'$ and taking expectations, we obtain

$$\begin{aligned} \mathrm{E}[\mathbf{Y}(t)\mathbf{Y}(t)'] &= \mathbf{B}\mathrm{E}[\mathbf{a}(t-1)\mathbf{Y}(t-1)'] && (94)\\ &= \mathbf{B}\mathrm{E}[\mathbf{a}(t-1)(\boldsymbol{\Phi}\mathbf{a}(t-1) + \boldsymbol{\nu}(t-1))'] && (95)\\ &= \mathbf{B}\mathbf{J}[\mathbf{C}_0^Z - \mathbf{V} - \mathbf{R}]\mathbf{J}'\boldsymbol{\Phi}', && (96) \end{aligned}$$

where we have exploited the independence relationships (12),(13), and (36). Now, letting

$$\mathbf{C}_1^Y \equiv \mathrm{E}[\mathbf{Y}(t)\mathbf{Y}(t-1)'], \tag{97}$$

and noting that $\mathbf{C}_1^Z = \mathbf{C}_1^Y$ as a consequence of (5), we can write

$$\mathbf{B} = \mathbf{C}_1^Z\mathbf{J}'(\mathbf{J}[\mathbf{C}_0^Z - \mathbf{V} - \mathbf{R}]\mathbf{J}')^{-1}. \tag{98}$$

Thus, we can obtain an estimate of $\mathbf{B}$ by substituting estimates of $\mathbf{C}_1^Z, \mathbf{C}_0^Z, \mathbf{V}$, and $\mathbf{R}$ into (98).

3.3 Estimation of $\mathbf{C}_0^Z, \mathbf{C}_1^Z$, and $\mathbf{C}_0^Y$

As shown above, in order to obtain estimates of the Kalman-filter model parameters, we need estimates of the covariance matrices $\mathbf{C}_0^Z, \mathbf{C}_1^Z$, and $\mathbf{C}_0^Y$. We estimate these matrices by MOM. In particular, let

$$\hat{c}_0(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \equiv \frac{1}{T}\sum_{t=1}^{T}(Z(\mathbf{s}_i;t) - \hat{\mu}_Z)(Z(\mathbf{s}_j;t) - \hat{\mu}_Z), \tag{99}$$

where

$$\hat{\mu}_Z \equiv \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} Z(\mathbf{s}_i;t), \tag{100}$$

is the estimate of the "grand" mean associated with the Z process. Then,

$$\hat{\mathbf{C}}_0^Z \equiv [\hat{c}_0(Z(\mathbf{s}_i), Z(\mathbf{s}_j))]_{i,j=1,\ldots,n} \ . \tag{101}$$

Similarly, the (i,j)-th element of the lag-one covariance matrix is estimated by

$$\hat{c}_1(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \equiv \frac{1}{T-1}\sum_{t=2}^{T}(Z(\mathbf{s}_i;t) - \hat{\mu}_Z)(Z(\mathbf{s}_j;t-1) - \hat{\mu}_Z), \tag{102}$$

and then

$$\hat{\mathbf{C}}_1^Z \equiv [\hat{c}_1(Z(\mathbf{s}_i), Z(\mathbf{s}_j))]_{i,j=1,\ldots,n} \ . \tag{103}$$

Finally, given an estimate of $\mathbf{R}$ (47) and $\mathbf{C}_0^Z$, we obtain the covariance matrix estimate of the Y process:

$$\hat{\mathbf{C}}_0^Y \equiv \hat{\mathbf{C}}_0^Z - \hat{\mathbf{R}}. \tag{104}$$

4 Selection of Model Basis Functions

As stated previously, we can choose any set of basis functions $\{\phi_k(\cdot)\}$, as long as they are complete, orthonormal, and are defined at any location $\mathbf{s}$ in domain D. Thus, there are many possible choices for these functions. For instance, we could choose eigenfunctions based on our *a priori* physical understanding of the system of interest. In a multivariate context, the theoretical eigenfunctions of the complete set of hydrodynamical equations which govern atmospheric motions could, in principle, be selected. More realistically, one could choose the eigenfunctions of some simplified form of the hydrodynamical equations of motion (e.g., the shallow water equations or perhaps the quasigeostrophic system). In a univariate context, one could choose physical basis functions based on general partial differential equations which can be related to certain kinds of physical phenomena (e.g., the eigenfunctions of the physically based covariance models of Whittle (1954) or Vecchia (1985)). Or, if we know little about the governing dynamics of a phenomenon of interest, we could choose some general basis set, either

empirical (e.g., EOFs) or specified (e.g., orthogonal polynomials, wavelets). Theoretically, it makes no difference which approach we take. However, certain choices are more advantageous from a practical standpoint. As discussed below, we focus our presentation on the EOF basis set.

4.1 The EOF Basis Set

In two spatial dimensions, we chose the EOF basis because it has a long history of use in the atmospheric sciences and more importantly, as will be discussed below, because it has certain optimality properties with regard to truncation. The use of EOFs in spatial prediction has been considered by Cohen and Jones (1969), Creutin and Obled (1982), Obled and Creutin (1986), Shriver and O'Brien (1995), and Reynolds et al. (1996). Creutin and Obled (1982) point out that the EOF approach to spatial prediction naturally accounts for anisotropic and heterogeneous covariance structure. For a recent discussion of EOFs and spatial prediction, see Guttorp and Sampson (1994). A comprehensive overview of EOFs, related to their use as a summarization tool in diagnostic studies of large atmospheric data sets, can be found in Preisendorfer (1988). As discussed by Buell (1972), when data are evenly distributed in space (i.e., gridded), EOF analysis is essentially equivalent to principal component analysis as defined in multivariate statistics.

Because we are assuming a spatially continuous observation and state process, we must obtain the EOF basis through a Karhunen-Loève (K-L) expansion (e.g., see Papoulis 1965, p.457-461). Given some spatio-temporal process $X(\mathbf{s};t)$ with $\mathbf{s} \in D$, $t \in \{1,2,\ldots\}$, suppose,

$$\mathrm{E}[X(\mathbf{s};t)] = 0, \tag{105}$$

and define the covariance matrix as

$$E[X(\mathbf{s};t)X(\mathbf{r};t)] \equiv c_0^X(\mathbf{s},\mathbf{r}), \tag{106}$$

which need not be stationary in space, but is assumed to be invariant in time. The K-L

expansion allows the covariance function to be decomposed as follows:

$$c_0^X(\mathbf{s},\mathbf{r}) = \sum_{k=1}^{\infty} \lambda_k \psi_k(\mathbf{s})\psi_k(\mathbf{r}), \tag{107}$$

where $\{\psi_k(\cdot) : k = 1,\ldots,\infty\}$ are the eigenvectors and $\{\lambda_k : k = 1,\ldots,\infty\}$ are the associated eigenvalues of the Fredholm integral equation

$$\int_D c_0^X(\mathbf{s},\mathbf{r})\psi_k(\mathbf{s})\, d\mathbf{s} = \lambda_k \psi_k(\mathbf{r}), \tag{108}$$

and

$$\int_D \psi_k(\mathbf{s})\psi_l(\mathbf{s}) d\mathbf{s} = \begin{cases} 1 & \text{for } k = l, \\ 0 & \text{otherwise.} \end{cases} \tag{109}$$

Assuming completeness, we can then expand $X(\mathbf{s};t)$ according to

$$X(\mathbf{s};t) = \sum_{k=1}^{\infty} f_k(t)\psi_k(\mathbf{s}), \tag{110}$$

where we call $\{\psi_k(\mathbf{s}) : \mathbf{s} \in D\}$ the k-th EOF and often refer to the associated time series $f_k(t)$ as the k-th principal component time series, or "amplitude" time series. This time series is derived from the projection of the X process onto the EOF basis,

$$f_k(t) = \int_D X(\mathbf{s};t)\psi_k(\mathbf{s})\, d\mathbf{s}, \tag{111}$$

and it is easy to verify that these time series are uncorrelated; that is,

$$\mathrm{E}[f_i(t) f_k(t)] = \delta_{ik}\lambda_k, \tag{112}$$

where δ_{ik} is one when $i = k$, and zero otherwise.

If we truncate the expansion (110) at J,

$$X_J(\mathbf{s};t) \equiv \sum_{k=1}^{J} f_k(t)\psi_k(\mathbf{s}), \tag{113}$$

then it can be shown (e.g., Freiberger and Grenander 1965; Davis 1976) that the EOF decomposition minimizes the variance of the truncation error, $\mathrm{E}\{[X(\mathbf{s};t) - X_J(\mathbf{s};t)]^2\}$, and is thus optimal in this regard when compared to all other basis sets.

4.2 Implementation of the EOF Basis

Data are always discrete. Therefore, in practice we must solve numerically the Fredholm integral equation (108) to obtain the EOF basis functions. Cohen and Jones (1969) and Buell (1972, 1975) give numerical quadrature solutions to this problem. The numerical quadrature approaches for discretizing the integral equation succeed in that they give estimates for the eigenfunctions and eigenvalues that are weighted in some manner according to the spatial distribution of the data locations, but only for the eigenfunctions at locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ for which there are data. In our case, the eigenfunctions are assumed to be non-stochastic. Thus, we could apply some relatively simple interpolation scheme to the eigenfunctions (obtained through quadrature) to obtain estimates at locations for which we do not have data. Similar to the "finite element" approaches for numerical integration used in engineering, Obled and Creutin (1986) demonstrate elegantly that this interpolation procedure can be linked to the numerical discretization by means of a set of canonical basis functions. We follow the approach outlined in Obled and Creutin (1986).

We define a canonical basis set as

$$\{e_i(\mathbf{s}) : i = 1, \ldots, n\}, \tag{114}$$

such that

$$e_i(\mathbf{s}_l) = \delta_{il}\,; \quad i, l = 1, \ldots, n, \tag{115}$$

where recall that δ_{il} is one when $i = l$ and zero, otherwise. In other words, the function $e_i(\mathbf{s}_l), i = 1, \ldots, n$, must equal one at location $\mathbf{s}_l$, and zero at every other data location. It is clear then that these basis functions have very local support. In two spatial dimensions, possible choices for basis functions include piecewise constant functions (i.e., equal to one in a neighborhood of $\mathbf{s}_l$ defined by an appropriate polygon region), facetlike linear functions (i.e., equal to one at $\mathbf{s}_l$ and linearly decreasing to zero at each neighboring data location), and thin-plate spline functions (i.e., equal to one at $\mathbf{s}_l$ and smoothly decreasing to zero at neighboring locations, but in a non-linear manner). In fact, these three basis sets are spline functions of increasing order, related to their "smoothness" (see, e.g., Chui 1988).

We selected the facetlike linear basis set for our discretization scheme. In two dimensions this requires a fixed triangulation of the observation network. One such triangulation is the well-known Delaunay triangulation (e.g., see Cressie 1993, p.373-374), which has received much attention because when it is used for planar interpolation, the greatest distances over which interpolations must be carried out are smaller than for any other triangulation. Given such a triangulation, consider an arbitrary triangle $\triangle_{ijk}$ in the network with vertices at $\{\mathbf{s}_i, \mathbf{s}_j, \mathbf{s}_k\}$. It is straightforward to show (e.g., Mori 1983, Obled and Creutin 1986) that if we are given some location $\mathbf{s}$ in $\triangle_{ijk}$ (or on its edge), then we obtain

$$e_i(\mathbf{s}) = \frac{(y_j - y_k)x - (x_j - x_k)y + (x_j y_k - x_k y_j)}{2A_i} \tag{116}$$

$$e_j(\mathbf{s}) = \frac{(y_k - y_i)x - (x_k - x_i)y + (x_k y_i - x_i y_k)}{2A_j} \tag{117}$$

$$e_k(\mathbf{s}) = \frac{(y_i - y_j)x - (x_i - x_j)y + (x_i y_j - x_j y_i)}{2A_k}, \tag{118}$$

where

$$A_i \equiv (1/2)[(x_j - x_i)(y_k - y_i) - (x_k - x_i)(y_j - y_i)] \tag{119}$$

$$A_j \equiv (1/2)[(x_k - x_j)(y_i - y_j) - (x_i - x_j)(y_k - y_j)] \tag{120}$$

$$A_k \equiv (1/2)[(x_i - x_k)(y_j - y_k) - (x_j - x_k)(y_i - y_k)], \tag{121}$$

and where we have assumed that, in Cartesian coordinates, the locations are represented by $\mathbf{s}_i = (x_i, y_i)$, $\mathbf{s}_j = (x_j, y_j)$, $\mathbf{s}_k = (x_k, y_k)$, and $\mathbf{s} = (x, y)$. Note that $|A_i| = |A_j| = |A_k|$, which is equal to the area of triangle $\triangle_{ijk}$.

Now, suppose we have some spatio-temporal process $X(\mathbf{s}; t)$ that we observe at discrete spatial locations $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and at time $t \in \{1, 2, \ldots\}$, and we would like to find a basis-function interpolation of X at location $\mathbf{s}_0$ and time t. We define this interpolated process according to

$$\tilde{X}(\mathbf{s}_0; t) \equiv \sum_{i=1}^{n} X(\mathbf{s}_i; t) e_i(\mathbf{s}_0). \tag{122}$$

Note that if the piecewise constant canonical basis set is chosen, this is equivalent to letting $\tilde{X}(\mathbf{s}_0; t)$ equal the value of $X(\cdot; t)$ at the measurement location nearest to $\mathbf{s}_0$ (i.e., the Thiessen

or Voronoi polygon method; see, e.g., Cressie 1993, p.374-375). If facetlike linear functions are chosen, this is equivalent to the Delaunay triangulation interpolation method (e.g., Cressie 1993, p.373-374).

Then the covariance function of this interpolated process is, for $\mathbf{s}, \mathbf{r} \in D$, $t \in \{1, 2, \ldots\}$,

$$\begin{aligned} c_0^{\tilde{X}}(\mathbf{s}, \mathbf{r}) &\equiv \mathrm{E}[\tilde{X}(\mathbf{s}; t)\tilde{X}(\mathbf{r}; t)] && (123) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} E[X(\mathbf{s}_i; t)X(\mathbf{s}_j; t)]e_i(\mathbf{s})e_j(\mathbf{r}) && (124) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_0^X(\mathbf{s}_i, \mathbf{s}_j)e_i(\mathbf{s})e_j(\mathbf{r}), && (125) \end{aligned}$$

where we have assumed temporal invariance and that X (and thus, $\tilde{X}$) has zero mean. We can then perform a K-L expansion on the interpolated process to obtain,

$$\int_D c_0^{\tilde{X}}(\mathbf{s}, \mathbf{r})\tilde{\psi}_k(\mathbf{r})\, d\mathbf{r} = \tilde{\lambda}_k\tilde{\psi}_k(\mathbf{s}), \tag{126}$$

where the eigenfunctions $\{\tilde{\psi} : k = 1, 2, \ldots\}$ are also linear combinations of $\{e_i(\mathbf{s}) : i = 1, \ldots, n\}$, namely,

$$\tilde{\psi}_k(\mathbf{s}) = \sum_{i=1}^{n} g_{ik}e_i(\mathbf{s}), \tag{127}$$

for $k = \{1, \ldots\}$. Now, from (115) we see that

$$g_{ik} = \tilde{\psi}_k(\mathbf{s}_i), \quad \text{for } i = 1, \ldots, n; k = 1, \ldots . \tag{128}$$

That is, $\{g_{ik}\}$ are effectively the eigenfunction values of the interpolated process at locations where we have data. They differ from $\{\psi_k(\mathbf{s}_i)\}$ due to the discretization. Now, upon substitution of (125) and (128) into (126), we obtain for $k = 1, \ldots, n$:

$$\begin{aligned} \int_D \sum_{i=1}^{n}\sum_{j=1}^{n} c_0^X(\mathbf{s}_i, \mathbf{s}_j)e_i(\mathbf{s})e_l(\mathbf{r}) \sum_{l=1}^{n} g_{lk}e_l(\mathbf{r})\, d\mathbf{r} &= \tilde{\lambda}_k \sum_{i=1}^{n} g_{ik}e_i(\mathbf{s}) && (129) \\ \sum_{i=1}^{n} e_i(\mathbf{s}) \left\{ \sum_{j=1}^{n}\sum_{l=1}^{n} c_0^X(\mathbf{s}_i, \mathbf{s}_j)g_{lk} \int_D e_j(\mathbf{r})e_l(\mathbf{r})\, d\mathbf{r} \right\} &= \tilde{\lambda}_k \sum_{i=1}^{n} g_{ik}e_i(\mathbf{s}) && (130) \\ \sum_{i=1}^{n} e_i(\mathbf{s}) \left\{ \sum_{j=1}^{n}\sum_{l=1}^{n} c_0^X(\mathbf{s}_i, \mathbf{s}_j)\, E_{il}\, g_{lk} \right\} &= \tilde{\lambda}_k \sum_{i=1}^{n} g_{ik}e_i(\mathbf{s}), && (131) \end{aligned}$$

where

$$E_{jl} \equiv \int_D e_j(\mathbf{r}) e_l(\mathbf{r})\, d\mathbf{r}. \tag{132}$$

Term-by-term comparison of the coefficients of $e_i(\mathbf{s}), i = 1, \ldots, n$ yields n equations of the form

$$\sum_{j=1}^{n}\sum_{l=1}^{n} c_0^X(\mathbf{s}_i, \mathbf{s}_j) E_{jl} g_{lk} = g_{ik}\tilde{\lambda}_k, \tag{133}$$

which we can write in matrix notation as

$$\mathbf{C}_0^X \mathbf{E}\mathbf{G} = \mathbf{G}\tilde{\mathbf{\Lambda}}, \tag{134}$$

where the matrices are given by

$$\mathbf{C}_0^X \equiv [c_0^x(\mathbf{s}_i, \mathbf{s}_j)]_{i,j=1,\ldots,n} \tag{135}$$

$$\mathbf{E} \equiv [E_{jl}]_{j,l=1,\ldots,n} \tag{136}$$

$$\mathbf{G} \equiv [g_{lk}]_{l,k=1,\ldots,n} \tag{137}$$

$$\tilde{\mathbf{\Lambda}} \equiv [\mathrm{diag}(\tilde{\lambda}_k)]_{k=1,\ldots,n}\ . \tag{138}$$

Given the covariance matrix $\mathbf{C}_0^X$ and the E matrix, we can solve the eigensystem (134) for G and $\tilde{\mathbf{\Lambda}}$.

As we have noted, the elements of G correspond to the eigenfunctions at each measurement location, considering the effect of the discretization. Then the eigenfunction values $\tilde{\psi}_k(\mathbf{s}), k = 1, \ldots, n$, at *any* location $\mathbf{s}$ are found using the interpolation formula (127). As shown by Obled and Creutin (1986), these interpolated eigenfunctions satisfy the orthonormality requirement since

$$\int_D \tilde{\psi}_k(\mathbf{r})\tilde{\psi}_l(\mathbf{r})\, d\mathbf{r} = \int_D \sum_{i=1}^{n} g_{ik} e_i(\mathbf{r}) \sum_{j=1}^{n} g_{jl} e_j(\mathbf{r})\, d\mathbf{r} \tag{139}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ik} E_{ij} g_{jl} \tag{140}$$

$$= \delta_{kl}. \tag{141}$$

Therefore, in order to solve the eigensystem (134) we must determine the elements of the matrix $\mathbf{E}$. Given an arbitrary triangle Δ_{ijk} from a specified triangulation of the observation

network, it can be shown (e.g., Mori 1983; Obled and Creutin 1986) that for the facetlike linear function canonical basis set, the contribution to the element E_{ii} corresponding to triangle Δ_{ijk} is given by

$$\int_{\Delta_{ijk}} e_i(\mathbf{s})e_i(\mathbf{s})\, d\mathbf{s} = (1/6)|\Delta_{ijk}|\,, \tag{142}$$

where $|\Delta_{ijk}|$ is the area of the triangle Δ_{ijk}. The contribution to the element $E_{ij}(i \neq j)$ corresponding to triangle Δ_{ijk} is

$$\int_{\Delta_{ijk}} e_i(\mathbf{s})e_j(\mathbf{s})\, d\mathbf{s} = (1/12)|\Delta_{ijk}|. \tag{143}$$

Contributions corresponding to triangles that do not have $\mathbf{s}_i$ or $\mathbf{s}_j$ as one of its vertices are zero. Thus, an algorithm for generating the $\mathbf{E}$ matrix goes as follows:

- Label each observation location with an integer from the set $\{1, 2, \ldots, n\}$, corresponding to positions in the $n \times n$ matrix $\mathbf{E}$.
- Initially set each element of $\mathbf{E}$ to zero.
- Find a triangulation of the observation location network (e.g., with a Delaunay triangulation).
- Calculate the area of each triangle in the network. For a given triangle Δ_{ijk} with $i, j, k \in \{1, 2, \ldots, n\}$, refer to this area as $|\Delta_{ijk}|$.
- For each triangle Δ_{ijk}, add $|\Delta_{ijk}|/6$ to the diagonal elements E_{ii}, E_{jj}, andE_{kk} and add $|\Delta_{ijk}|/12$ to the nondiagonal elements $E_{ij}, E_{ji}, E_{ik}, E_{ki}, E_{jk}$, and E_{kj}.

Thus, we need only a triangulation and the areas of the associated triangles to obtain $\mathbf{E}$.

For the model (1),(2),(3), we are interested in the eigenfunctions of the Y_K process (3), which we obtain through the Y process (2). Then, after obtaining the matrix $\mathbf{E}$ for the observation network, we use the estimate of $\mathbf{C}_0^Y$ (104) in place of $\mathbf{C}_0^X$ in (134) and solve for $\mathbf{G}$ and $\tilde{\mathbf{\Lambda}}$. The matrix $\mathbf{\Phi}$ in (28) is then set equal to $\mathbf{G}$. We can then interpolate the elements of $\mathbf{\Phi}$ to locations where we do not have data by using (127).

5 Simulation Example

This section describes the results obtained from the implementation of the spatially descriptive, temporally dynamic (SDTD) model presented in the preceding sections. In particular, the model is applied to a simulated spatio-temporal data set. In addition, the results from the SDTD model are compared to the those obtained from a simple kriging analysis applied separately to the spatial field at each time (i.e., assuming no dynamical structure in time).

5.1 Description of Simulation

For the simulation, we assume a separable spatio-temporal model consisting of a first-order Markov process in time, with spatially colored noise:

$$X(\mathbf{s};t) = \gamma X(\mathbf{s};t-1) + \eta(\mathbf{s};t), \tag{144}$$

where γ is an autoregressive parameter (assumed to be 0.75 in the simulation), and $\eta(\mathbf{s};t)$ is a spatially colored noise process. We let $\mathbf{s}$ be defined at grid nodes on a 17 by 17 grid with uniform grid spacing (equal to 0.125 in both the x- and y-directions), as shown in Figure 1a. In addition, we let t take integer values from 1 to 150. We then assume that this process X is contaminated with measurement error according to the measurement equation

$$Z(\mathbf{s};t) = X(\mathbf{s};t) + \epsilon(\mathbf{s};t), \tag{145}$$

where $\epsilon(\mathbf{s};t)$ is assumed to be Gaussian white noise with zero mean and variance $\sigma_\epsilon^2 = 0.015$.

The colored noise process $\eta(\mathbf{s};t)$ is found at each time t by generating a grid of Gaussian random deviates, on which we apply a windowed spatial moving average. Specifically, $\eta(\mathbf{s};t)$ is the average of all the Gaussian random deviates within a specified radius r of the location $\mathbf{s}$. Clearly, as this radius increases, there is more spatial dependence. This method of calculating the spatial noise is easier to implement than the matrix decomposition methods described, for example, in Cressie (1993; Section 3.6.1). The covariance function corresponding to this moving average approach is isotropic, stationary, and very similar in structure to the spherical covariance function (Cressie 1993, p.61). In this simulation, we choose the radius r to be 0.625,

which gives very strong spatial dependence. Biases near the edge of the simulation grid due to the truncation of the moving window are accounted for by defining a larger grid for the determination of η than used in the simulation.

The simulation is performed so that the process is known at all locations on the grid, and at all times. Then, 40 observation locations are chosen at random, with the exception that no locations are allowed within a .75 by .75 square centered in the middle of the grid. This "hole" will be one of the areas at which we will be interested in obtaining predictions. The "observation" locations are shown in Figure 1b. Thus, the simulated values at these locations are considered to be the "data".

As described in Section 4.2, a triangulation of the observation location network is needed for the EOF basis-set implementation of the SDTD model. The Delaunay triangulation of the data locations, as obtained from an algorithm by S. Fortune (Fortune 1987), is shown in Figure 1b.

Our goal is to predict at spatial locations where we do not have data. Predictions are made for three different sections of the simulation grid: the entire grid (referred to hereafter as grid "A"), a grid covering the data "hole" in the center of the grid (referred to as grid "B"; see Figure 1a), and a cross-section in the x-direction at location zero in the y-direction (see Figure 1a).

To implement the SDTD model, we must choose a truncation parameter K and measurement error estimate $\hat{\sigma}_\epsilon^2$. The truncation parameter was chosen to be seven, based on an examination of the eigenvalues of the estimated matrix $\hat{\mathbf{C}}_0^Y$ [e.g., see (77)]. Then, based on a plot of empirical covariances of the Z-process (Figure 2), we selected the estimated measurement error $\hat{\sigma}_\epsilon^2$ to be 0.011. A discussion of the choice of different values for the estimated measurement error is given below.

As described in Section 3.1.2, estimates of the ν-process covariance matrix $\hat{\mathbf{V}}$ were found and a model was fitted. Figure 3 shows the estimated ν covariances at different lags, the "bin" averages of those covariances, and a model fit. Note the negative covariances in the region near

a spatial lag of 0.5. This structure is common in meteorological variables. Thiébaux (1976) discusses several covariance models that can be used in this setting. Of these, we chose the so-called "cosine-modified Gaussian" model,

$$c_\nu(h; a, b, g, c_\nu(0)) = [b\cos(ah) + c_\nu(0) - b]e^{-gh^2}, \tag{146}$$

where a, b, g, and $c_\nu(0)$ are model parameters ($c_\nu(0)$ is the variance at lag zero). In our case, we specify $c_\nu(0)$ based on (83) and then estimate the remaining parameters via the weighted least squares approach (81). Figure 3 shows the fitted model. As noted in Section 3.1.2, we have the option of using the modeled covariances according to (146) or simply assuming that ν is white noise. Although the analyses presented here were performed under both assumptions, only the results for the white noise case are presented. The results for both approaches were similar, but objective validation statistics (see Section 5.2.1 below) showed that the white noise assumption leads to marginally better predictions with these data.

5.2 SDTD Model Results with Simulated Data

To examine the temporal characteristics of the SDTD predictions, we compare the predicted time series at a given location $\hat{X}(\mathbf{s}; \cdot)$ with the true signal $X(\mathbf{s}; \cdot)$ and the noisy signal $Z(\mathbf{s}; \cdot)$. Figure 4a shows this comparison at an observation location (-.5, .125), and Figure 4b shows the comparison at a location (0,0) where we do not have data. The Kalman filter predictions at both locations do a good job of tracking the true signal and filtering the measurement error. However, in typical Kalman filter fashion, they both have some difficulty predicting the extreme peaks and valleys of the series.

Although informative, the time series comparison given above does not provide information on the model's ability to predict in space. A visually powerful means of considering both space and time is to examine a spatial cross section as it varies in time (which we refer to as an *x-t plot*). Figure 5 shows a x-t contour plot for the noisy signal (Figure 5a), the true signal (Figure 5b), and the predicted signal (Figure 5c) given the noisy data. The cross-section was taken at $(y = 0)$ as shown in Figure 1a. The SDTD model has done an excellent job of capturing the

essence of the spatio-temporal signal from relatively noisy observations. This achievement is even more striking when we consider that only three of the 17 spatial locations in this cross-section were observation locations. However, there are differences between the predicted and the true signals. Not surprisingly, the predicted signal is smoother than the true signal. In addition, there are times when the predicted signal is too strong (e.g., $t = 40$) and others where it is too weak (e.g., $t = 110$).

We now consider the SDTD model's ability to predict in two spatial dimensions and in time. Unfortunately, without the aid of computer animation, such a depiction is difficult. However, in Figure 6 we show surface plots of the predictions on the full grid (grid A) "stacked" in time. The left column of Figure 6 shows the true signal $X(\cdot\,;t)$ for $t = 75$ to $t = 78$. The right column of Figure 6 shows the corresponding prediction from the SDTD model. Clearly, the predictions are much smoother than the true fields. However, the predictions have done an excellent job of capturing the "spatial trends" in the true signal. In addition, the prediction has captured the "valley" in the true signal relatively well, with the exception of time $t = 78$, which does not properly capture the location of the valley. In general, the model does seem to be capturing the dynamical nature of the underlying process.

The square root of the mean-squared prediction errors (MSPEs) (see Eqn. 63) for this simulation are given in Figure 7. Note that since there are no missing observations, the MSPE is constant across time. As is the case with kriging predictors, the MSPE is lowest around data locations. Typically, we would expect the MSPE to be significantly larger in data sparse areas such as the "hole" in the center and at the edges of our simulation grid. However, the SDTD model appears to be doing a good job in this region since the MSPEs are not significantly larger than at nearby measurement locations.

5.2.1 Objective Validation Statistics

The preceding discussion of the performance of the SDTD model predictions has been qualitative. For a quantitative measure of the model's precision and accuracy, we consider the

following three validation statistics:

$$CR_1(t) = (1/m)\sum_{j=1}^{m}\{(X(\mathbf{s}_j;t) - \hat{X}(\mathbf{s}_j;t))/\hat{\sigma}_X(\mathbf{s}_j;t)\}, \tag{147}$$

$$CR_2(t) = [(1/m)\sum_{j=1}^{m}\{(X(\mathbf{s}_j;t) - \hat{X}(\mathbf{s}_j;t))/\hat{\sigma}_X(\mathbf{s}_j;t)\}^2]^{1/2}, \tag{148}$$

and

$$CR_3(t) = [(1/m)\sum_{j=1}^{m}\{(X(\mathbf{s}_j;t) - \hat{X}(\mathbf{s}_j;t))^2]^{1/2}, \tag{149}$$

where $\hat{X}(\mathbf{s}_j;t)$ is the prediction of the process X at location $\mathbf{s}_j$ and time t, and where $\mathbf{s}_j$ is defined at each location in the prediction grid. Furthermore, $\hat{\sigma}_X(\mathbf{s}_j;t)$ is the corresponding model-estimated standard deviation of the MSPE. These validation statistics are modified from the corresponding cross-validation statistics used by Carroll and Cressie (1996) and outlined in Cressie (1993, p.102). As noted in those sources, $CR_1(t)$ provides an estimate of the unbiasedness of the predictors for each time and should be very close to 0; $CR_2(t)$ gives a measure of the accuracy of the mean squared prediction errors and should be very close to 1; and $CR_3(t)$ is a check of the "goodness of prediction", similar to the PRESS statistic used in the evaluation of multiple regression analysis (e.g., Draper and Smith 1981). One would like $CR_3(t)$ to be small so that the predicted values are close to the true values. One can then look at the time series of these statistics. For an "average" view of each, one could also look at the time mean of these statistics. We will take the latter approach in this presentation.

As shown in Table 1, for the case of the SDTD model analysis as presented above, the $\bar{CR}_1$,$\bar{CR}_2$, and $\bar{CR}_3$ statistics (where the "bar" denotes the time mean) are .016, 1.024, and .087, respectively, for grid B (i.e., covering the "hole" region). As a comparison, these values are .015, 1.134, and .089, respectively, for predictions on grid A (i.e., the entire simulation grid). This is somewhat surprising since we would expect the predictions over the "hole" region to be worse since there are no observations there. Thus, an apparent strength of the SDTD model and the EOF implementation is that predictions can be made reasonably well over such large gaps in data.

The SDTD analysis was run with several different values of $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\nu^2$, as well as different values of the truncation parameter K. The values selected for this presentation (i.e., $\hat{\sigma}_\epsilon^2 = .011$ and $\hat{\sigma}_\nu^2 = .005$) were chosen based on an *ad hoc* comparison of the values of the validation statistics for different scenarios. For instance, such a comparison provided the justification for using the white noise representation for the covariance function of the ν process (83), rather than the model (146). However, the selection of model parameters by this approach is made more difficult since we have little guidance on how to partition the two error variances σ_ϵ^2 and σ_ν^2 (i.e., they are non-identifiable). Further effort should be directed at finding optimal partitions.

5.3 Comparison to Other Methods

The validation statistics outlined above provide an objective way in which to compare the accuracy and precision of two or more different prediction approaches. A simple approach to spatial prediction given spatio-temporal data is to perform some sort of optimal spatial average (i.e., kriging) independently at each time. Since we are assuming that we know our simulated process has zero mean, we could use the optimal estimation procedure known in geostatistics as *simple kriging* (e.g., Cressie 1993, p.110). Since our simulated data include measurement error, we must use simple kriging in the presence of such error. Such a formulation is not known to exist in the literature. Consequently, we present the method in Appendix B. Simple kriging can then be used as a benchmark with which to compare the SDTD results on the simulated data. The choice of simple kriging for this comparison is further warranted by the fact that the SDTD model prediction equation (53) has a simple kriging-like component to account for the residual (ν) noise process, as discussed in Section 2.1.

5.3.1 Comparison of Simple Kriging and the SDTD Model

The measurement error for the simple kriging analysis was chosen to be the same value (i.e., $\hat{\sigma}_\epsilon^2 = 0.011$) as that used in the SDTD analysis reported above. Choosing the same value for each analysis makes for a better comparison of the two methods. Figure 2 shows the empirical

covariance estimates of the data Z, as well as the fitted covariance model. A spherical model was chosen for lags higher than .15, and an exponential model was selected for lags 0 to 0.15. An average of the time-averaged variances at each data location was chosen as the estimate of the covariance at lag 0 (equal to 0.0354 in this analysis). As a simplification, the same covariance model is applied at all values of t. Although this assumption seems severe, Carroll and Cressie (1996) show that such an assumption works quite well for the prediction of snow water equivalent.

Table 1 shows the validation statistics for both the SDTD and kriging predictors. For grid B (the central "hole" region), the SDTD predictions are clearly less biased (from $\bar{CR}_1$), and the overall predictive fit (from $\bar{CR}_3$) is better than for the simple kriging case. The ratio of the $\bar{CR}_3$ statistics between the two methods shows that the SDTD approach is approximately 10 % more accurate in its predictions than the simple kriging approach. The kriging method is slightly better at capturing the empirical MSPE via the modeled MSPE (from $\bar{CR}_2$), although both are quite good. For grid A (the entire region), the SDTD analysis is less biased, better able to capture the MSPE, and gives a better fit (by approximately 15 %) than the simple kriging analysis. Thus, the SDTD model is performing noticeably better than the simple kriging approach.

Table 1 Validation statistics for the SDTD model and simple kriging (SK) analyses on simulated data

Method	Grid	$\bar{CR}_1$	$\bar{CR}_2$	$\bar{CR}_3$
SDTD	A	0.0148	1.1337	0.0885
SK	A	0.0343	1.4251	0.1053
SDTD	B	0.0160	1.0235	0.0870
SK	B	0.0948	0.9995	0.0964

The standard deviation of the MSPE for the simple kriging analysis is given in Figure 8. There are several interesting features of this plot as compared to the standard deviation of the MSPE for the SDTD analysis (Figure 7). In particular, the prediction errors are significantly lower around the measurement locations in the kriging analysis than in the SDTD analysis.

However, the SDTD analysis has significantly lower MSPE values in the data "hole" at the grid center, and in the region near the edges of the simulation grid. These are regions where kriging traditionally performs poorly. Thus, it is clear that the SDTD method has improved on traditional kriging predictors in areas where data coverage is less.

6 South China Sea Precipitation Example

In this section, the SDTD model is applied to a "real world" example. In particular, we have chosen to apply this model to a precipitation data set covering the South China Sea (SCS) region of southwestern Asia. Precipitation was chosen because it is a variable for which reliable spatial predictions are typically difficult to obtain. This difficulty arises because precipitation variability encompasses such a wide range of spatial and temporal scales, and because the physical characterization of the processes leading to these scales of variability and their interactions are relatively poorly understood. Specifically, we selected the SCS region because it is representative of the interaction of these scales of variability, and because this region includes effects from both midlatitude and tropical weather systems (e.g., Chen and Chen 1995). In addition, the lack of precipitation data over the SCS provides a good test of the SDTD model's ability to predict in regions with sparse spatial sampling.

6.1 Data

Monthly precipitation data were obtained from the National Climatic Data Center (NCDC) operated by the National Oceanic and Atmospheric Administration (NOAA). The data were from the first version of the Global Historical Climatology Network (GHCN) surface baseline data set. This data set was created by a joint effort of the NCDC and the Carbon Dioxide Information Analysis Center at the Oak Ridge National Laboratory in 1992 for the purposes of providing a data set to be used to monitor and detect climate change (Vose et al. 1992). Over 7,500 precipitation stations are included in the GHCN, with data extending back into the 1600's. All of the data have a quality control flag that gives an indication of possible serial

and spatial continuity problems.

We selected 135 stations around the SCS for the period 1959 to 1988, as shown in Figure 9. With the exception of three stations, all of the stations chosen for this analysis had high quality data for at least 85% of the time periods between January 1959 and December 1988. The three stations on the coast of Vietnam were chosen for the analysis, even though they did not meet this 85% criterion (the data were not available after 1975). These stations were included because it was thought that their presence on the western coast of the SCS would help define the EOF basis functions used in the analysis.

The goal of this analysis is then to predict the monthly precipitation over the SCS at locations and times where there are no data. We have selected a prediction grid that covers the majority of the SCS and that has uniform grid point separations of 2° in latitude and longitude. This prediction grid is shown in Figure 9.

6.2 Exploratory Data Analysis

Precipitation amounts can vary greatly from one geographical region to another. Thus, we expect that precipitation variability may be a function of the mean precipitation amount. This is indeed the case with our data set as presented in Figure 10. This figure shows a plot of the natural log of the mean versus the natural log of the variance of precipitation (time-averaged at each observation location). The linear relationship evident in this plot is quite striking for "real" data. An ordinary least squares (OLS) fit, weighted by the number of observations used to calculate the respective means and variances, is shown in Figure 10 as well. The slope of this OLS best-fit line is 1.548 and the intercept is 1.099. This linear relationship (in logs) can easily be shown to be equivalent to a power of the mean model given by:

$$\hat{\sigma}^2(\mathbf{s}_i) = 3.000\ \hat{\mu}(\mathbf{s}_i)^{1.548}, \tag{150}$$

where $\hat{\sigma}^2(\mathbf{s}_i)$ and $\hat{\mu}(\mathbf{s}_i)$ are the estimated variance and mean at data location $\mathbf{s}_i$, respectively. The fact that the variance is a function of the mean invalidates the homogeneity of variance assumptions inherent in the SDTD model development. Therefore, we must perform a variance

stabilizing transformation such that transformed data have constant variance (e.g., Snedecor and Cochran 1989, p.286-287). This procedure is outlined below.

6.2.1 Variance Stabilizing Transformation and Data Preparation

Assume we are given data X such that the variance is dependent on the mean in the following sense:

$$\sigma^2(\mu) = k\,\mu^{\beta}, \tag{151}$$

where $\sigma^2(\mu)$ is the the variance of X, μ is the mean of X, and k and β are unknown parameters. We then seek the transformation $f(X)$ that gives constant variance. It can easily be shown that expanding $f(X)$ in a Taylor series about μ and requiring $\text{var}(f(X))$ to be constant leads to the following relationship:

$$f(X) \propto X^{1-\beta/2}. \tag{152}$$

In our case, we have estimated β to be 1.548, and let k equal 1.0, so the appropriate variance stabilizing transformation is

$$f(X) = X^{0.226}, \tag{153}$$

which is very close to the familiar 4-th root transformation.

In practice, after applying the transformation (153) to the precipitation data, we must remove the seasonal mean effect. This is done by calculating the time mean for each of the 12 months of the year, for each station. Then, the appropriate monthly mean is subtracted from the truncated data. These means must also be estimated at locations where we do not have data if we are to predict at such locations. We can obtain these estimates by applying the Delaunay triangulation machinery developed for estimating the basis functions at locations where we do not have data (see Section 4.2).

After running the Kalman filter implementation of the SDTD model, we must transform the predictions and MSPEs back to the original scale. It is well known that such a transformation induces a bias in the prediction (e.g., Cressie 1993, p.135-138). Therefore, we must correct for this bias. In our case, recall that we have defined the Kalman filter predictor in (53) to

be $\hat{Y}(\mathbf{s};t\,|\,t)$. Also, in the derivation of the SDTD model we assumed our data were given by the Z process, which is taken to have zero mean and homogeneous variance. Now, we further assume that these data Z were obtained through a variance stabilizing transformation on the original data X [i.e., $Z = f(X)$] and that we are interested in predicting the smooth process W on the original scale, corresponding to smooth process Y on the transformed scale. That is, we would like to obtain $\hat{W}(\mathbf{s};t\,|\,t)$ from the Kalman filter prediction $\hat{Y}(\mathbf{s};t\,|\,t)$. Analogous to Cressie (1993, p.137) we can use Taylor series expansions to obtain the following approximate relationship for the unbiased predictor:

$$\hat{W}(\mathbf{s};t\,|\,t) \approx \tilde{f}(\hat{Y}(\mathbf{s};t\,|\,t)) + (1/2)\tilde{f}''(\hat{\mu}_Y(\mathbf{s};t))\hat{\sigma}^2_Y(\mathbf{s};t), \tag{154}$$

where $\tilde{f}(\ ) \equiv f^{-1}(\ )$, $\hat{\mu}_Y(\mathbf{s};t)$ is the estimate of the seasonal mean of the transformed data, and $\hat{\sigma}^2_Y(\mathbf{s};t)$ is the MSPE of the Y process as given by (63). It is also easy to show that the MSPE for the W process is given approximately by

$$\hat{\sigma}^2_W(\mathbf{s};t) \approx \{\tilde{f}'(\hat{\mu}_Y(\mathbf{s};t))\}^2\hat{\sigma}^2_Y(\mathbf{s};t). \tag{155}$$

In our case, the function $f(\cdot)$ is defined by (153), and so the function $\tilde{f}(\cdot)$ is defined as the inverse of (153), namely $\tilde{f}(Z) = Z^{1/0.226}$.

6.3 Implementing the SDTD Model

In order to implement the SDTD model, we must acquire a triangulation of the observation network. First, we transform our map coordinates to an equal area projection. In particular, we use the cylindrical equal area projection described in Pearson (1990, p.129-132). This transformation is needed because the discretization of the continuous EOF analysis depends on weights that are based on the areas of the triangles in the network, as described in Section 4.2. Thus, we do not want a map projection with non-equal areas, which would bias our determination of the appropriate weights. The Delaunay triangulation on this equal area projection, as determined by the algorithm of Fortune (1987), is shown in Figure 11.

We expect that the data are contaminated by measurement error since they are collected and disseminated by such a large variety of jurisdictions. To examine this, we plot the estimated

empirical semi-variogram in Figure 12. As is evident in this plot, there is a "nugget effect" associated with these data, thus implying microscale variability and/or measurement error. We assume that the entire nugget effect is due to measurement error, and choose a value of $\hat{\sigma}_\epsilon^2 = 0.03$.

Next, we examine the eigenvalues of the estimated matrix $\hat{\mathbf{C}}_0^Y$ to help determine the truncation value K. We selected $K = 25$ since the first 25 eigenvalues imply that the estimated matrix $\hat{\mathbf{C}}_0^{Y_K}$ would then account for approximately 90% of the variance of $\hat{\mathbf{C}}_0^Y$. Sensitivity analyses show that the results are relatively robust to the choice of this parameter, so long as it is reasonably large.

As described in Section 3.1.2, estimates of the ν-process covariance matrix $\hat{\mathbf{V}}$ were found. Figure 13 shows the estimated ν covariances at different lags, as well as the "bin" averages of those covariances. We note the large spike at lag zero, and the steep decline at higher lags. Also note the negative covariance feature near lags of 500km, which was also present in the simulated data (Figure 3). Based on the fact that the SDTD predictions for the simulated data showed better results under the assumption of a white noise covariance structure for the ν process, we took that approach here as well. Then, we estimated $\hat{\sigma}_\nu^2 = 0.07$ by the approach suggested in Section 3.1.2 [see (83)].

6.4 SDTD Model Results with Precipitation Data

The SDTD model was run with the transformed precipitation data described in Section 6.2.1 and the predictions were transformed back to the original scale for presentation. Figure 14 shows time series plots of the predictions at three locations for the 10 year period from January 1979 to December 1988: (a) a location on the southern coast of China near (114°E, 22°N); (b) a location in the middle of the SCS at (114°E, 12°N); and (c) a location along the northwestern coast of Borneo near (114°E, 4°N). In addition to the predicted precipitation (solid line), these figures show the monthly mean precipitation (dotted line), and Figures 14a and 14c show the observed precipitation (dashed line). In each case, the predictions appear to have captured the appropriate structure in the time variability of precipitation. Although

some of the extreme peaks are clearly missed by the predicted time series (e.g., the extreme event in 1982 in Figure 14a), in general, the predictions are able to capture the "direction" of the deviation from the seasonal mean. Although not shown, the Kalman filter successfully predicts missing values in the time series.

We now consider the prediction over the grid shown in Figure 1 to cover the SCS. For illustration, we present the results for 1979. This year was very active meteorologically in the region, with a strong SCS monsoon, Mei-Yu front, and 30-60 day oscillation (e.g., see Chen and Chen 1995). Figure 15 shows surface plots of the predicted precipitation (left column) and the associated (square) root MSPE (right column) for March, April, and May of 1979. Note that in March, the precipitation shows two peaks, one in the north and one in the south, with a minimum over the SCS. The root MSPE plot corresponds similarly to these peaks and valleys as expected since (155) shows that the transformation of the MSPE is dependent on the mean. Thus, areas with larger mean precipitation will have larger MSPEs, and areas with small means will have smaller MSPEs. The northern peak corresponds to the convection associated with the Mei-Yu front and the southern peak corresponds to that associated with the Intertropical Convergence Zone (ITCZ). Figures 16 and 17 show the surface plots of predicted precipitation and MSPE standard deviations for the summer (June, July, August) and fall (September, October,November), respectively. Notice that throughout the spring (March, April, May) months, the central SCS shows a relative minimum in predicted precipitation. By July, the ITCZ has begun to move northward, leading to larger precipitation amounts over the central SCS. After reaching its maximum intensity in August, the ITCZ begins to retreat to the south. By November, the ITCZ is firmly entrenched in the southern region, and the north is very dry. This migration is also apparent in satellite-derived precipitation estimates for the same time period (e.g., Chen and Chen 1995). Thus, the SDTD model appears to capture the dynamic evolution of the precipitation over the data-sparse SCS.

7 Conclusion

In this investigation we have presented a new spatio-temporal statistical model that considers the influence of both spatial and temporal variability. The model is temporally dynamic in that it exploits the unidirectional flow of time in an autoregressive framework, and is spatially descriptive in that no causative interpretation is associated with its spatially colored noise. The inclusion of a measurement equation naturally leads to the development of a spatio-temporal Kalman filter. The Kalman filter implementation allows us to predict in time and in space, and to account for missing data.

We demonstrated this method by applying it to a simulated spatio-temporal data set. The model was shown to capture the temporal dynamic structure as well as the spatial structure of the simulated data, although a certain amount of smoothing in the predictions was evident. A comparison with the predictions obtained from a simple kriging analysis applied separately to each time showed that, for this simulation, the proposed approach was generally superior in its predictive skill.

In addition, the model was applied to 30 years of monthly precipitation data from the South China Sea region of Asia. The model seems to capture the dynamic evolution of the spatial processes associated with the precipitation in this region.

There is much that can be done to further the development of this approach. In particular, we should conduct additional simulations to test the performance of this model under different conditions. For instance, it is suspected that as the temporal dependence decreases, the proposed model may not perform any better than the simple kriging approach. Additional study of the effect of different partitions of the measurement error and the error associated with the ν process would also be useful.

An investigation of the application of different basis functions should be performed. It was noted in Section 4 that any complete and orthonormal basis set can be used in the model implementation. Other than the EOF basis used here, potential basis sets include orthogonal polynomials and wavelets. We have implemented a Legendre polynomial basis set in a simula-

tion of spatio-temporal data with one spatial dimension. The results are encouraging, and the method is in many ways more pleasing since the triangulation and EOF discretization are not needed. Wavelets have high potential in this regard as they have compact support and fast algorithms exist for their implementation.

We also note that the evaluation of this model was concerned with spatial prediction, rather than temporal prediction. Clearly, the ability to predict in space and time is a major strength of this model, one that is not at all prevalent in the literature. This has great potential for meteorological investigations where a "first-guess" prediction can be useful in numerical weather prediction. The statistical approach presented here could be used for the first-guess of variables which are not modeled well by short-term integrations of the numerical weather prediction models (e.g., water vapor and precipitation). We suspect that for such temporal predictions to be useful, the model should be extended to include additional time lags. Such an extension should be straightforward, mainly taking into account additional lag covariance matrices.

It would also be useful to allow the measurement locations to vary in time. In this case, for example, we could perform spatial prediction of sea surface temperatures that are obtained from transient ocean ships, as well as stationary buoy observations.

In addition, we could explore the possibility of including additional variables (say, precipitation and temperature together) in the prediction. Such a multivariate approach should be a straightforward extension of the current model. In particular, the state-space formulation of the model easily can be easily modified to allow a multivariate approach.

We could also include explanatory variables to aid in the prediction. For instance, if we wanted to predict precipitation over the central United States, we would probably want to consider the effect of sea surface temperatures in the Pacific. So, although the state variable in our model would still be precipitation, we would include sea surface temperature as a predictor, but it would not be predicted. This approach should be a straightforward extension of the current model.

Finally, from a statistical standpoint, perhaps the most challenging endeavor would be to implement this model in a hierarchical Bayesian framework. Given that the Kalman filter can be interpreted from a Bayesian perspective, the estimation of model parameters given the data, as applied in this presentation, corresponds to an empirical Bayes approach. Alternatively, we could go the additional step and assign prior distributions to the model parameters. This is the hierarchical Bayesian approach. We must then decide what kind of prior information we should use. One option would be to specify some diffuse prior, thereby reflecting our potential ignorance of the prior distribution. A more physically interesting approach would be to make use of the governing system of equations of the climate system (if they are known). In other words, we recognize that although climate models do not give anywhere near a complete description of the climate system, they do, by definition, operate under the same basic physical laws as the "true" climate system. So, perhaps we could use a climate simulation to obtain our prior knowledge about certain parameters in the spatio-temporal statistical model. We could then employ the hierarchical Bayesian approach with these priors and obtain the corresponding posterior distributions. The study of the similarities and differences between these prior and the posterior distributions is a potentially rich source of physical information about the climate system.

Acknowledgements

This research was sponsored by the U.S. Department of Energy, Office of Energy Research, Environmental Sciences Division, Office of Health and Environmental Research, under the first author's appointment to the Graduate Fellowships for Global Change administered by Oak Ridge Institute for Science and Education. The second author's research was sponsored in part by a co-operative agreement between the U.S. Environmental Protection Agency and Iowa State University.

Appendix A: The Spatio-Temporal Kalman Filter

This section outlines the derivation of the spatio-temporal Kalman filter. The approach used here follows a univariate derivation given in Meinhold and Singpurwalla (1983). We assume normality throughout.

Recall from (35) and (32) that the measurement and state equations are given by

$$Z(\mathbf{s};t) = \boldsymbol{\phi}(\boldsymbol{s})'\mathbf{a}(t) + \nu(\mathbf{s};t) + \epsilon(\mathbf{s};t) \quad \text{(A. 1)}$$

$$\mathbf{a}(t) = \mathbf{H}\mathbf{a}(t-1) + \boldsymbol{J}\boldsymbol{\eta}(t), \quad \text{(A. 2)}$$

where we recall that

$$\mathbf{R} \equiv \text{var}[\epsilon(t)] \quad \text{(A. 3)}$$

$$\mathbf{V} \equiv \text{var}[\boldsymbol{\nu}(t)] \quad \text{(A. 4)}$$

$$\mathbf{Q} \equiv \text{var}[\boldsymbol{\eta}(t)]. \quad \text{(A. 5)}$$

We are interested in the joint distribution of $(\mathbf{a}(t)', \mathbf{Z}(t)')$ conditional on $\mathbf{Z}^*(t-1)$, where

$$\mathbf{Z}^*(t-1) \equiv [\mathbf{Z}(t-1), \ldots, \mathbf{Z}(1)]. \quad \text{(A. 6)}$$

Consider Bayes' theorem:

$$P(\mathbf{a}(t) \,|\, \mathbf{Z}(t), \mathbf{Z}^*(t-1)) \propto P(\mathbf{Z}(t) \,|\, \mathbf{a}(t), \mathbf{Z}^*(t-1))\, P(\mathbf{a}(t) \,|\, \mathbf{Z}^*(t-1)), \quad \text{(A. 7)}$$

where the first term on the right-hand side is the "likelihood" and the second term on the right-hand side is the "prior" distribution. At time $t-1$, our knowledge about $\mathbf{a}(t-1)$ is given by:

$$(\mathbf{a}(t-1) \,|\, \mathbf{Z}^*(t-1)) \sim N(\hat{\mathbf{a}}(t-1 \mid t-1), \mathbf{P}(t-1 \mid t-1)), \quad \text{(A. 8)}$$

where

$$\hat{\mathbf{a}}(t-1 \mid t-1) \equiv \text{E}[\mathbf{a}(t-1) \mid \mathbf{Z}^*(t-1)] \quad \text{(A. 9)}$$

$$\mathbf{P}(t-1 \,|\, t-1) \equiv \text{var}[\mathbf{a}(t-1) \mid \mathbf{Z}^*(t-1)]. \quad \text{(A. 10)}$$

Prior to observation $\mathbf{Z}(t)$, the "best" prediction of $\mathbf{a}(t)$ is given by the state equation (A.2). Thus, we obtain

$$\begin{aligned}
\hat{\mathbf{a}}(t \mid t-1) &= \mathrm{E}[\mathbf{a}(t) \mid \mathbf{Z}^*(t-1)] && \text{(A. 11)}\\
&= \mathrm{E}[\mathbf{H}\mathbf{a}(t-1) + \mathbf{J}\boldsymbol{\eta}(t) \mid \mathbf{Z}^*(t-1)] && \text{(A. 12)}\\
&= \mathrm{E}[\mathbf{H}\mathbf{a}(t-1) \mid \mathbf{Z}^*(t-1)] && \text{(A. 13)}\\
&= \mathbf{H}\hat{\mathbf{a}}(t-1 \mid t-1), && \text{(A. 14)}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{P}(t \mid t-1) &= \mathrm{var}[\mathbf{a}(t) \mid \mathbf{Z}^*(t-1)] && \text{(A. 15)}\\
&= \mathrm{var}[\mathbf{H}\mathbf{a}(t-1) + \mathbf{J}\boldsymbol{\eta}(t) \mid \mathbf{Z}^*(t-1)] && \text{(A. 16)}\\
&= \mathbf{H}\mathbf{P}(t-1 \mid t-1)\mathbf{H}' + \mathrm{var}(\mathbf{J}\boldsymbol{\eta}(t)) && \text{(A. 17)}\\
&= \mathbf{H}\mathbf{P}(t-1 \mid t-1)\mathbf{H}' + \mathbf{J}\mathbf{Q}\mathbf{J}'. && \text{(A. 18)}
\end{aligned}$$

Note that we have used the fact that $\boldsymbol{\eta}(t)$ is uncorrelated with $\mathbf{Z}^*(t-1)$. Then, the "prior" distribution is given by

$$(\mathbf{a}(t) \mid \mathbf{Z}^*(t-1)) \sim N(\hat{\mathbf{a}}(t \mid t-1), \mathbf{P}(t \mid t-1)). \qquad \text{(A. 19)}$$

Now, we assume that $\mathbf{Z}(t)$ has been observed. Then, we want to determine the posterior distribution from (A.7). First, we need to find the "likelihood" distribution $P(\mathbf{Z}(t) \mid \mathbf{a}(t), \mathbf{Z}^*(t-1))$. We define the one-step prediction error for $\mathbf{Z}(t)$ as:

$$\mathbf{e}(t) \equiv \mathbf{Z}(t) - \hat{\mathbf{Z}}(t \mid t-1), \qquad \text{(A. 20)}$$

where

$$\begin{aligned}
\hat{\mathbf{Z}}(t \mid t-1) &\equiv \mathrm{E}[\mathbf{Z}(t) \mid \mathbf{Z}^*(t-1)] && \text{(A. 21)}\\
&= \mathrm{E}[\boldsymbol{\Phi}\mathbf{a}(t) + \boldsymbol{\nu}(t) + \boldsymbol{\epsilon}(t) \mid \mathbf{Z}^*(t-1)] && \text{(A. 22)}\\
&= \mathrm{E}[\boldsymbol{\Phi}\mathbf{a}(t) \mid \mathbf{Z}^*(t-1)] && \text{(A. 23)}\\
&= \boldsymbol{\Phi}\hat{\mathbf{a}}(t \mid t-1), && \text{(A. 24)}
\end{aligned}$$

and where we have used the fact that $\boldsymbol{\nu}(t)$ and $\boldsymbol{\epsilon}(t)$ are uncorrelated with $\mathbf{Z}^*(t-1)$. Thus,

$$\mathbf{e}(t) = \mathbf{Z}(t) - \boldsymbol{\Phi}\hat{\mathbf{a}}(t \mid t-1). \tag{A. 25}$$

If $\boldsymbol{\Phi}$, $\mathbf{H}$, and $\hat{\mathbf{a}}(t-1 \mid t-1)$ are assumed known, then observing $\mathbf{Z}(t)$ is equivalent to observing $\mathbf{e}(t)$. Now, we use the measurement equation (A.1) to write

$$\mathbf{e}(t) = \boldsymbol{\Phi}(\mathbf{a}(t) - \hat{\mathbf{a}}(t \mid t-1)) + \boldsymbol{\nu}(t) + \boldsymbol{\epsilon}(t). \tag{A. 26}$$

Thus,

$$\begin{aligned} \mathrm{E}[\mathbf{e}(t) \mid \mathbf{a}(t), \mathbf{Z}^*(t-1)] &= \mathrm{E}[\boldsymbol{\Phi}(\mathbf{a}(t) - \hat{\mathbf{a}}(t \mid t-1)) \mid \mathbf{a}(t), \mathbf{Z}^*(t-1)] && \text{(A. 27)} \\ &= \boldsymbol{\Phi}(\mathbf{a}(t) - \hat{\mathbf{a}}(t \mid t-1)), && \text{(A. 28)} \end{aligned}$$

where we have used the fact that $\boldsymbol{\nu}(t)$ and $\boldsymbol{\epsilon}(t)$ are both uncorrelated with $\mathbf{a}(t)$ and $\mathbf{Z}^*(t-1)$. Then, using (A.3) and (A.4) we obtain the distributional relationship:

$$(\mathbf{e}(t) \mid \mathbf{a}(t), \mathbf{Z}^*(t-1)) \sim N(\boldsymbol{\Phi}(\mathbf{a}(t) - \hat{\mathbf{a}}(t \mid t-1)), \mathbf{R} + \mathbf{V}). \tag{A. 29}$$

We can now obtain the posterior distribution by making use of certain properties of the multivariate normal distribution. It is well known that if $\mathbf{e}(t)$ and $\mathbf{a}(t)$ are jointly multivariate normal, then we can write the conditional distribution as (e.g., see Johnson and Wichern 1992, p.138):

$$(\mathbf{e}(t) \mid \mathbf{a}(t), \mathbf{Z}^*(t-1)) \sim N(\boldsymbol{\mu}_e + \boldsymbol{\Sigma}_{ea}\boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{a}(t) - \boldsymbol{\mu}_a), \boldsymbol{\Sigma}_{ee} - \boldsymbol{\Sigma}_{ea}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ae}), \tag{A. 30}$$

where, in general, $\boldsymbol{\mu}_u$ is the mean of $\mathbf{u}$, and $\boldsymbol{\Sigma}_{uv}$ is the covariance between $\mathbf{u}$ and $\mathbf{v}$. Note, in our case,

$$\begin{aligned} \boldsymbol{\mu}_e &\equiv \mathrm{E}[\mathbf{e}(t) \mid \mathbf{a}(t), \mathbf{Z}^*(t-1)] && \text{(A. 31)} \\ &= \mathrm{E}[\mathbf{Z}(t) - \hat{\mathbf{Z}}(t \mid t-1) \mid \mathbf{a}(t), \mathbf{Z}^*(t-1)] && \text{(A. 32)} \\ &= \mathrm{E}[\mathbf{Z}(t) \mid \mathbf{Z}^*(t-1)] - \hat{\mathbf{Z}}(t \mid t-1) && \text{(A. 33)} \\ &= \mathbf{0}. && \text{(A. 34)} \end{aligned}$$

It can then be shown that

$$\Sigma_{ea} = \Phi \mathbf{P}(t \mid t-1), \tag{A. 35}$$

where we make use of the orthogonality of the predictor and the residual, as is well known from linear projection theory. Then, it can be easily shown that

$$\begin{aligned} \Sigma_{ee} - \Sigma_{ea}\Sigma_{aa}^{-1}\Sigma_{ae} &= \Sigma_{ee} - \Phi \mathbf{P}(t \mid t-1)\Phi' && \text{(A. 36)} \\ &= \mathbf{R} + \mathbf{V}. && \text{(A. 37)} \end{aligned}$$

Therefore,

$$\Sigma_{ee} = \mathbf{R} + \mathbf{V} + \Phi \mathbf{P}(t \mid t-1)\Phi'. \tag{A. 38}$$

Once again, we can make use of the multivariate normal distribution and obtain the conditional distribution (i.e., the "posterior" distribution):

$$(\mathbf{a}(t) \mid \mathbf{e}(t), \mathbf{Z}^*(t-1)) \sim N(\hat{\mathbf{a}}(t \mid t), \mathbf{P}(t \mid t)), \tag{A. 39}$$

where

$$\begin{aligned} \hat{\mathbf{a}}(t \mid t) &= \hat{\mathbf{a}}(t \mid t-1) + \mathbf{K}(t)[\mathbf{Z}(t) - \Phi \hat{\mathbf{a}}(t \mid t-1)] && \text{(A. 40)} \\ \mathbf{P}(t \mid t) &= \mathbf{P}(t \mid t-1) - \mathbf{K}(t)\Phi \mathbf{P}(t \mid t-1), && \text{(A. 41)} \end{aligned}$$

and the Kalman gain $\mathbf{K}(t)$ is given by

$$\mathbf{K}(t) = \mathbf{P}(t \mid t-1)\Phi'[\mathbf{R} + \mathbf{V} + \Phi \mathbf{P}(t \mid t-1)\Phi']^{-1}. \tag{A. 42}$$

One-step ahead predictions are then given by the "prior" distribution (A.19) to be

$$\begin{aligned} \hat{\mathbf{a}}(t \mid t-1) &\equiv \mathrm{E}[\mathbf{a}(t) \mid \mathbf{Z}^*(t-1)] && \text{(A. 43)} \\ &= \mathbf{H}\hat{\mathbf{a}}(t-1 \mid t-1) && \text{(A. 44)} \\ \mathbf{P}(t \mid t-1) &\equiv \mathrm{var}[\mathbf{a}(t) \mid \mathbf{Z}^*(t-1)] && \text{(A. 45)} \\ &= \mathbf{H}\mathbf{P}(t-1 \mid t-1)\mathbf{H}' + \mathbf{J}\mathbf{Q}\mathbf{J}'. && \text{(A. 46)} \end{aligned}$$

Appendix B: Simple Kriging in the Presence of Measurement Error

This appendix shows the derivation of the simple kriging equations in the presence of measurement error. The setting is a purely spatial one.

Consider the noisy spatial process $Z(\cdot)$ with zero mean, such that

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{B. 1}$$

where $Y(\cdot)$ is a zero mean smooth spatial process and $\epsilon(\cdot)$ is a white noise error process. We seek a linear predictor of the smooth process Y at some location $\mathbf{s}_0$ of the form

$$\hat{Y}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i), \tag{B. 2}$$

where $\{\mathbf{s}_i : i = 1, \ldots, n\}$ are measurement locations, and $\lambda_i, i = 1, \ldots, n$ are unknown parameters.

First, consider the case where $\mathbf{s}_0 \neq \mathbf{s}_i;\ i = 1, \ldots, n$. Then, the mean squared prediction error is defined as

$$\begin{aligned}
\sigma^2_{SK}(\mathbf{s}_0) &\equiv \mathrm{E}[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) - Y(\mathbf{s}_0)]^2 && \text{(B. 3)} \\
&= E[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) - Z(\mathbf{s}_0) + \epsilon(\mathbf{s}_0)]^2 && \text{(B. 4)} \\
&= \boldsymbol{\lambda}' \mathbf{C}^Z \boldsymbol{\lambda} - 2\boldsymbol{\lambda}' \mathbf{c}(\mathbf{s}_0) + c(\mathbf{s}_0, \mathbf{s}_0) - \sigma^2_\epsilon, && \text{(B. 5)}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\lambda} &\equiv (\lambda_1, \ldots, \lambda_n)' && \text{(B. 6)} \\
c(\mathbf{s}, \mathbf{r}) &\equiv \mathrm{E}[Z(\mathbf{s}) Z(\mathbf{r})] && \text{(B. 7)} \\
\mathbf{c}(\mathbf{r}) &\equiv (c(\mathbf{r}, \mathbf{s}_1), \ldots, c(\mathbf{r}, \mathbf{s}_n))' && \text{(B. 8)} \\
\mathbf{C}^Z &\equiv [c(\mathbf{s}_i, \mathbf{s}_j)]_{i,j=1,\ldots,n} && \text{(B. 9)} \\
\sigma^2_\epsilon &= \mathrm{E}[\epsilon(\mathbf{s})\epsilon(\mathbf{s})]. && \text{(B. 10)}
\end{aligned}$$

We then seek to minimize the mean squared prediction error (B.5) with respect to the λ's. Thus, taking the derivative of (B.5) with respect to $\boldsymbol{\lambda}$ and equating the result with zero gives

$$\boldsymbol{\lambda}' = \mathbf{c}(\mathbf{s}_0)' [\mathbf{C}^Z]^{-1}. \tag{B. 11}$$

Plugging this into (B.5) then gives the mean squared prediction error:

$$\sigma^2_{SK}(\mathbf{s}_0) = c(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}(\mathbf{s}_0)'[\mathbf{C}^Z]^{-1}\mathbf{c}(\mathbf{s}_0) - \sigma^2_\epsilon \,. \tag{B. 12}$$

Now, consider the case where $\mathbf{s}_0 = \mathbf{s}_i$ for some i. For demonstration, let $\mathbf{s}_0 = \mathbf{s}_1$. Then, the mean squared prediction error is

$$\begin{aligned} \sigma^2_{SK}(\mathbf{s}_1) &= \mathrm{E}[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) - Y(\mathbf{s}_1)]^2 && \text{(B. 13)} \\ &= E[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) - Z(\mathbf{s}_1) + \epsilon(\mathbf{s}_1)]^2 && \text{(B. 14)} \\ &= \boldsymbol{\lambda}'\mathbf{C}^Z\boldsymbol{\lambda} - 2\boldsymbol{\lambda}'\tilde{\mathbf{c}}(\mathbf{s}_1) + c(\mathbf{s}_1, \mathbf{s}_1) - \sigma^2_\epsilon, && \text{(B. 15)} \end{aligned}$$

where

$$\tilde{\mathbf{c}}(\mathbf{s}_1) \equiv \mathbf{c}(\mathbf{s}_1) - (\sigma^2_\epsilon, 0, \ldots, 0)'. \tag{B. 16}$$

Minimizing the mean squared prediction error (B.15) as before gives

$$\boldsymbol{\lambda}' = \tilde{\mathbf{c}}(\mathbf{s}_1)'[\mathbf{C}^Z]^{-1} \,. \tag{B. 17}$$

Plugging this into (B.15) then gives the mean squared prediction error:

$$\sigma^2_{SK}(\mathbf{s}_1) = c(\mathbf{s}_1, \mathbf{s}_1) - \tilde{\mathbf{c}}(\mathbf{s}_1)'[\mathbf{C}^Z]^{-1}\tilde{\mathbf{c}}(\mathbf{s}_1) - \sigma^2_\epsilon \,. \tag{B. 18}$$

Thus, considering both cases, we can write the simple kriging equations in the presence of measurement error as

$$\begin{aligned} \boldsymbol{\lambda}' &= \mathbf{c}^*(\mathbf{s}_0)'[\mathbf{C}^Z]^{-1} && \text{(B. 19)} \\ \sigma^2_{SK}(\mathbf{s}_0) &= c(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}^*(\mathbf{s}_0)'[\mathbf{C}^Z]^{-1}\mathbf{c}^*(\mathbf{s}_0) - \sigma^2_\epsilon \,, && \text{(B. 20)} \end{aligned}$$

where

$$\mathbf{c}^*(\mathbf{s}_0) \equiv \begin{cases} \mathbf{c}(\mathbf{s}_0) & \text{if } \mathbf{s}_0 \neq \mathbf{s}_i \; i = 1, \ldots, n \\ \tilde{\mathbf{c}}(\mathbf{s}_0) & \text{if } \mathbf{s}_0 = \mathbf{s}_i \; i = 1, \ldots, n \,, \end{cases} \tag{B. 21}$$

and $\tilde{c}(\mathbf{s}_0)$ is given by (B.16). In general, $\tilde{\mathbf{c}}(\mathbf{s}_i)$ is simply equal to $\mathbf{c}(\mathbf{s}_i)$ with the i-th element replaced by $c(\mathbf{s}_i, \mathbf{s}_i) - \sigma^2_\epsilon$.

References

Bennett, R.J., 1979: *Spatial Time Series*, Pion Limited, 674pp.

Bilonick, R.A., 1983: Risk qualified maps of hydrogen ion concentration for the New York state area for 1966-1978. *Atmos. Environ.*,**17**, 2513-2524.

Bretherton, C.S., C. Smith, and J.M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, **5**, 541-560.

Buell, C.E., 1972: Integral equation representation for factor analysis. *J. Atmos. Sci.*, **28**, 1502-1505.

Buell, C.E., 1975: The topography of empirical orthogonal functions. *Preprints Fourth Conf. Prob. Stats. Atmos. Sci.*, American Meteorological Society, 188-193.

Carroll, S.S. and N. Cressie, 1996: A comparison of geostatistical methodologies used to estimate snow water equivalent. *Water Resources Bulletin.* Forthcoming.

Chen, T.-C. and J.-M. Chen, 1995: An observational study of the South China Sea monsoon during the 1979 summer: Onset and life cycle. *Mon. Wea. Rev.*, **123**, 2295-2318.

Chui, C.K., 1988: *Multivariate Splines*, Society for Industrial and Applied Mathematics, 189pp.

Cohen, A. and R.H. Jones, 1969: Regression on a random field. *J. Amer. Stat. Assoc.*, **64**, 1172-1182.

Cohn, S.E., and D.F. Parrish, 1991: The behavior of forecast error covariances for a Kalman filter in two dimensions. *Mon. Wea. Rev.*, **119**, 1757-1785.

Cressie, N., 1994: Comment on "An approach to statistical spatial-temporal modeling of meteorological fields" by M.S. Handcock and J.R. Wallis, *J. Amer. Stat. Assoc.*, **89**, 379-382.

Cressie, N.A.C, 1993: *Statistics for Spatial Data, Revised Edition,* Wiley, 900pp.

Creutin, J.D., and Ch. Obled, 1982: Objective analysis and mapping techniques for rainfall fields: An objective comparison. *Water. Res. Res.*, **18**, 413-431.

Daley, R., 1995: Estimating the wind field from chemical constituent observations: Experiments with a one-dimensional extended Kalman filter. *Mon. Wea. Rev.*, **123**, 181-198.

Daley, R., 1991: *Atmospheric Data Analysis*, Cambridge University Press, 457pp.

Davis, R.E., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific ocean. *J. Physical Oceanography,* **6**, 249-266.

Dee, D.P., 1995: On-line estimation of error covariance parameters for atmospheric data analysis. *Mon. Wea. Rev.*, **123**, 1128-1145.

Dee, D.P., 1991: Simplification of the Kalman filter for meteorological data assimilation. *Quart. J. Roy. Meteor. Soc.*, **117**, 365-384.

Dee, D.P., S. E. Cohn, A. Dalcher, and M. Ghil, 1985: An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Trans. Automatic Control*, **AC-30**, 1057-1065.

Draper, N.R. and H. Smith, 1981: *Applied Regression Analysis*, Second Edition, Wiley, 709pp. Eynon, B.P., and P. Switzer, 1983: The variability of rainfall acidity. *Canadian J. of Statist.*, **11**, 11-24.

Fortune, S.J., 1987: A sweepline algorithm for Voronoi diagrams. *Algorithmica*, **2**, 153-174.

Freiberger, W. and U. Grenander, 1965: On the formulation of statistical meteorology. *Review of the International Statistical Institute,* **33**, 59-86.

Gandin, L.S., 1963: *Objective Analysis of Meteorological Fields.* Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad (translated by Israel Program for Scientific Translations, Jerusalem, 1965).

Glahn, H.R., 1968: Canonical correlation and its relationship to discriminant analysis and multiple regression. *J. Atmos. Sci.*, **25**,23-31.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics,* Vol. 33, Academic Press, 141-266.

Ghil, M., S.E. Cohn, J. Tavantzis, K. Bube, and E. Isaacson, 1981: Applications of estimation theory to numerical weather prediction. *Dynamic meteorology: Data assimilation methods*, L. Bengtsson, M. Ghil, and E. Källén, eds., Springer-Verlag, 139-224.

Goodall, C., and K.V. Mardia, 1994: Challenges in multivariate spatio-temporal modeling. Proceedings of the XVIIth International Biometric Conference, Hamilton, Ontario Canada, 8-12 August 1994.

Guttorp, P., and P.D. Sampson, 1994: Methods for estimating heterogeneous spatial covariance functions with environmental applications. *Handbook of Statistics, Vol. 12*, G.P. Patil and C.R. Rao, eds., Elsevier Science, 661-689.

Haas, T.C., 1995: Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *J. Amer. Stat. Assoc.*, **90**, 1189-1199.

Haas, T.C., 1990a: Kriging and automated variogram modeling within a moving window. *Atmos. Environ.*, **24**A, 1759-1769.

Haas, T.C., 1990b: Lognormal and moving window methods of estimating acid deposition. *J. Amer. Stat. Assoc.*, **85**, 950-963.

Handcock, M.S., and J.R. Wallis, 1994: An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *J. Amer. Stat. Assoc.*, **89**, 368-390.

Haslett, J., 1989: Space time modelling in meteorology - A review. *Bull. Int. Statist. Int.*, **51**, 229-246.

Haslett, J., and A.E. Raftery, 1989: Spatio-Temporal modelling with long-memory dependence: Assessing Ireland's wind resource (with discussion). *J. Roy. Statist. Soc. Ser. C*, **38**, 1-50.

Hasselmann, K., 1988: PIPs and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns. *J. Geophys. Res.*,**93**, 11015 - 11021.

Host, G., H. Omre, and P. Switzer, 1995: Spatial interpolation errors for monitoring data. *J. Amer. Statist. Assoc.*, **90**, 853-861.

Huang, H.-C. and N. Cressie, 1996: Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis*, in press.

Hughes, J.P. and P. Guttorp, 1994a: A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resources Research*, **30**, 1535-1546.

Hughes, J.P. and P. Guttorp, 1994b: Incorporating spatial dependence and atmospheric data in a model of precipitation. *J. Appl. Meteor.*, **33**, 1503-1515.

Johnson, R.A. and D.W. Wichern, 1992: *Applied Multivariate Statistical Analysis*, Third Edition, Prentice Hall, 642pp.

Loader, C., and P. Switzer, 1992: Spatial covariance estimation for monitoring data. *Statistics in Environmental and Earth Sciences*, A. Walden and P. Guttorp, eds., Charles W. Griffin, 52-69.

Lorenz, E.N., 1956: Empirical orthogonal functions and statistical weather prediction. *Sci. Rept. No. 1, Statistical Forecasting Project*, MIT, 49 pp.

Mardia, K.V., and C. R. Goodall, 1993: Factorized models in spatial modeling. *Multivariate Environmental Statistics*, N.K. Bose, G.P. Patil and C.R. Rao, eds., North-Holland.

Matheron, G., 1963: Principles of geostatistics. *Economic Geology*, **58**, 1246-1266.

Meinhold, J. and N.D. Singpurwalla, 1983: Understanding the Kalman filter. *The American Statistician,* **37**, 123-127.

Miller, R.N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.*, **51**, 1037-1056.

Obled, Ch., and J.D. Creutin, 1986: Some developments in the use of empirical orthogonal functions for mapping meteorological fields. *J. Clim. Appl. Meteor.*, **25**, 1189-1204.

Oehlert, G.W., 1993: Regional trends in sulfate wet deposition.*J. Amer. Stat. Assoc.*, **88**, 390-399.

Papoulis, A., 1965: *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, Inc., pp. 457-461.

Pearson, F., 1990: *Map Projections: Theory and Applications*, CRC Press.

Preisendorfer, R.W., 1988: *Principal Component Analysis in Meteorology and Oceanography*, Elsevier, 425 pp.

Reynolds, R.W., R.E. Livezey, T.M. Smith, and D.C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions: Part I. *J. Climate,* in press.

Rouhani, S., and D.E. Meyers, 1990: Problems in spatio-temporal kriging of geohydrological data. *Math. Geol.*, **22**, 611-623.

Rouhani, S., and H. Wackernagel, 1990: Multivariate geostatistical approach to space-time data analysis. *Water Res. Res.,* **26**, 585-591.

Sampson, P.D., and P. Guttorp, 1992: Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Stat. Assoc.*, **87**, 108-119.

Shriver, J.F., and J.J. O'Brien, 1995: Low-frequency variability of the equatorial pacific ocean using a new pseudostress data set: 1930-1989. *J. Climate,* **8**, 2762-2786.

Stein, M., 1986: A simple model for spatial-temporal processes. *Water Res. Res.,* **22**, 2107-2110.

Thiebaux, H.J., and Pedder, M.A., 1987: *Spatial objective analysis with applications in atmospheric science.* Academic Press.

Vecchia, A.V., 1985: A general class of models for stationary two-dimensional random processes. *Biometrika*, **72**, 281-291.

von Storch, H., G. Bürger, R. Schnur, and J.-S. von Storch, 1995: Principal oscillation patterns: a review. *J. Climate,* **8**, 377-400.

von Storch, H., T. Bruns, I. Fischer-Bruns, and K. Hasselmann, 1988: Principal oscillation pattern analysis of the 30- to 60-day oscillation in a general circulation model equatorial troposphere. *J. Geophys. Res.*, **93**, 11022-11036.

Vose, R.S., R.L. Schmayer, P.M. Steurer, T.C. Peterson, R. Heim, T.R. Karl, and J.K. Eischeid, 1992: The Global Historical Climatology Network: Long-term monthly temperature, precipitation, sea level pressure, and station pressure data. NDP-041. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Whittle, P., 1954: On stationary processes in the plane. *Biometrika*, **41**, 434-449.

Zhang, Y., 1995: Autoregressive models for continuous space-time processes. Ph.D. Dissertation, University of Colorado, Denver, CO.

Zucchini, W. and P. Guttorp, 1991: A hidden Markov model for space-time precipitation. *Water Res. Res.*, **27**, 1917-1923.

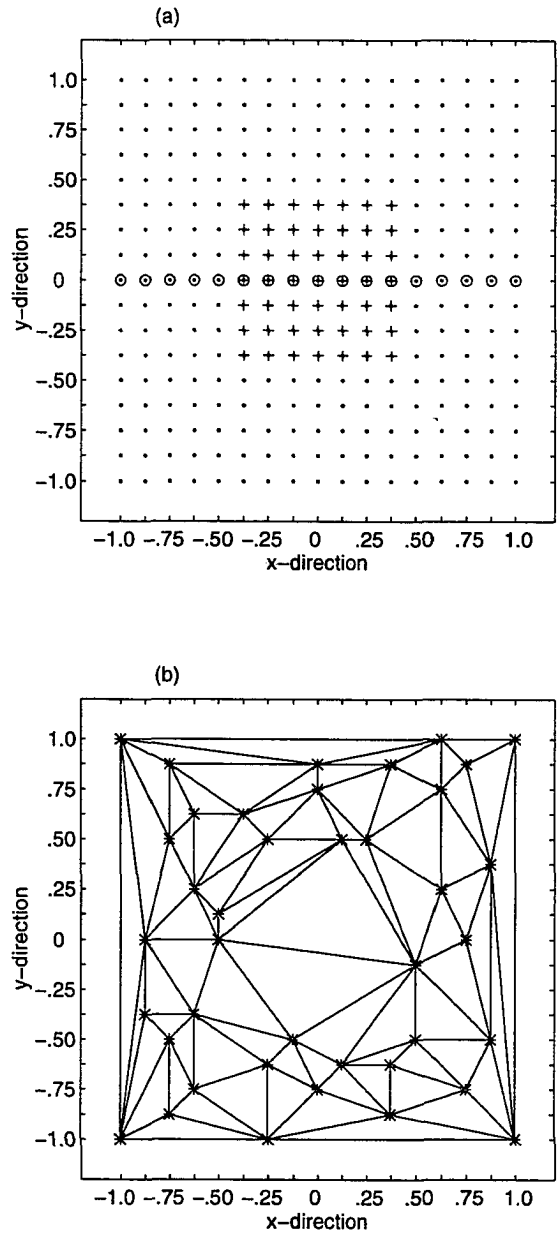

Figure 1 (a) Simulation grid "A" (·), prediction grid "B" (+), and prediction cross-section (o); (b) Observation locations for simulated data (∗) and Delaunay triangulation of observation network.

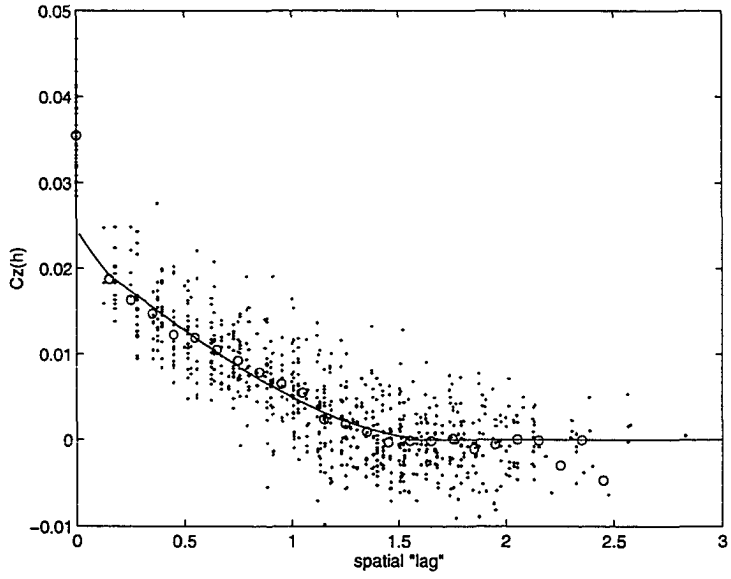

Figure 2 Estimated covariances of simulated Z process (+); "bin" averages (o); and the weighted least squares fitted covariance model, which is taken to be exponential from lag 0 to lag .125, and spherical from lag .125 to lag 3.

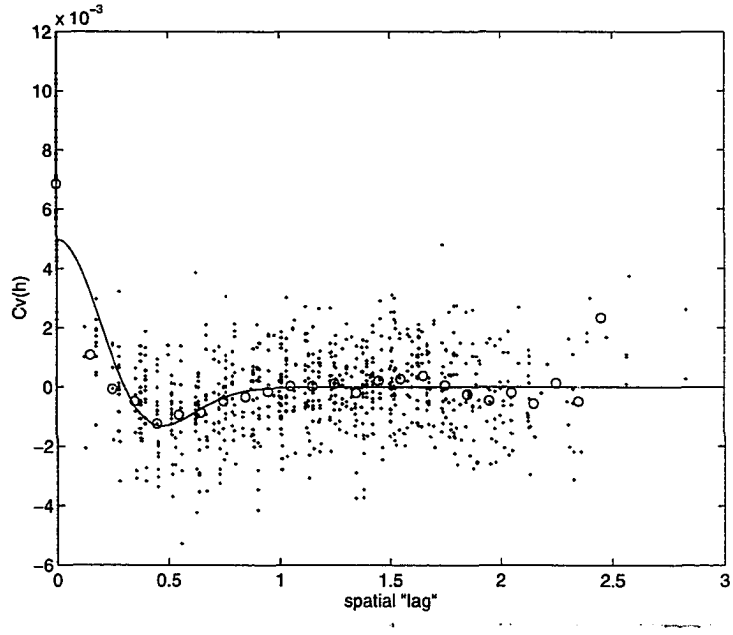

Figure 3 Estimated covariances of the ν process from the simulation (+); "bin" averages (o); and the weighted least squares fitted covariance model (Eqn. 146).

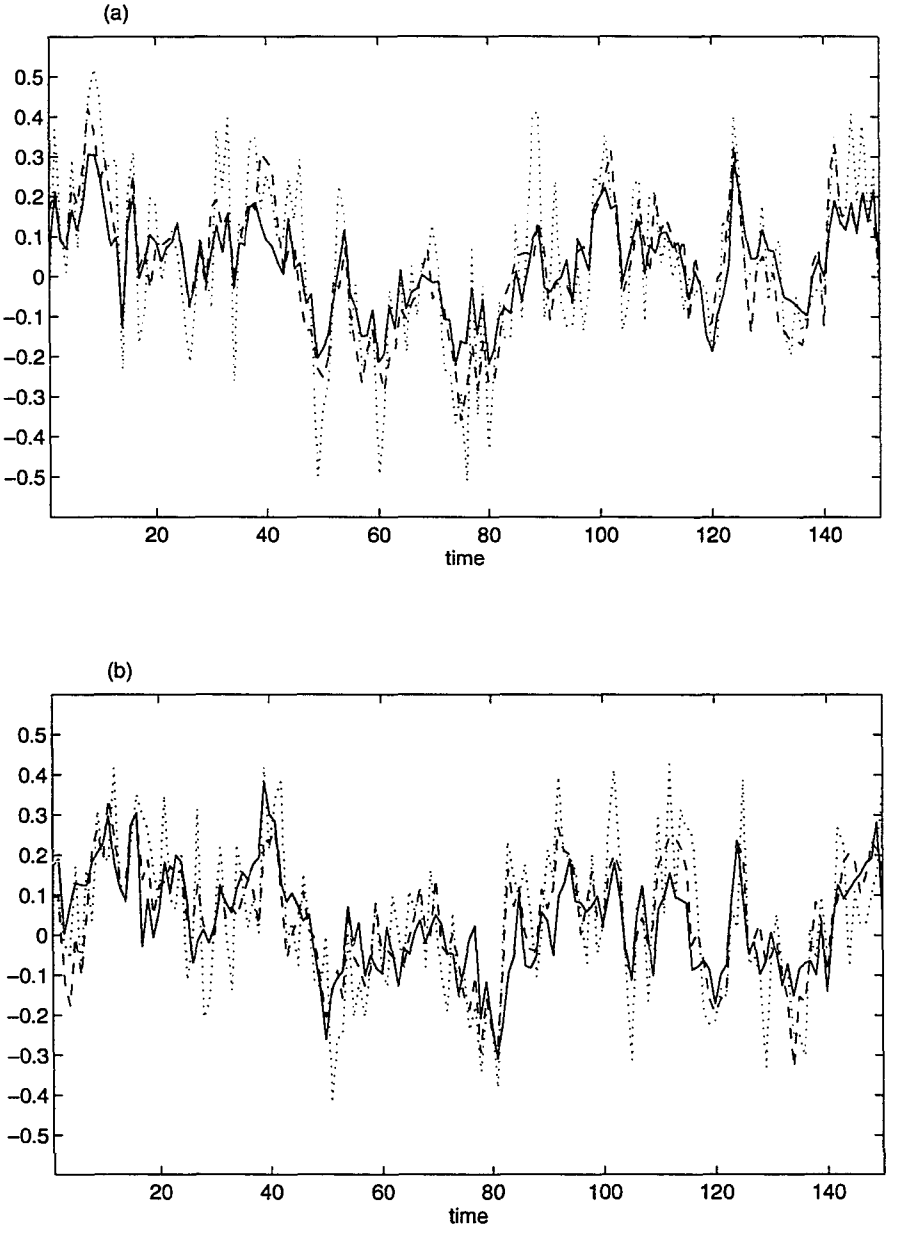

Figure 4 Predicted time series $\hat{Y}$ (solid line), true time series Y, and the associated noisy time series Z for the simulation at locations: (a) (-0.5, .125), a measurement location, and (b) (0,0), a non-measurement location.

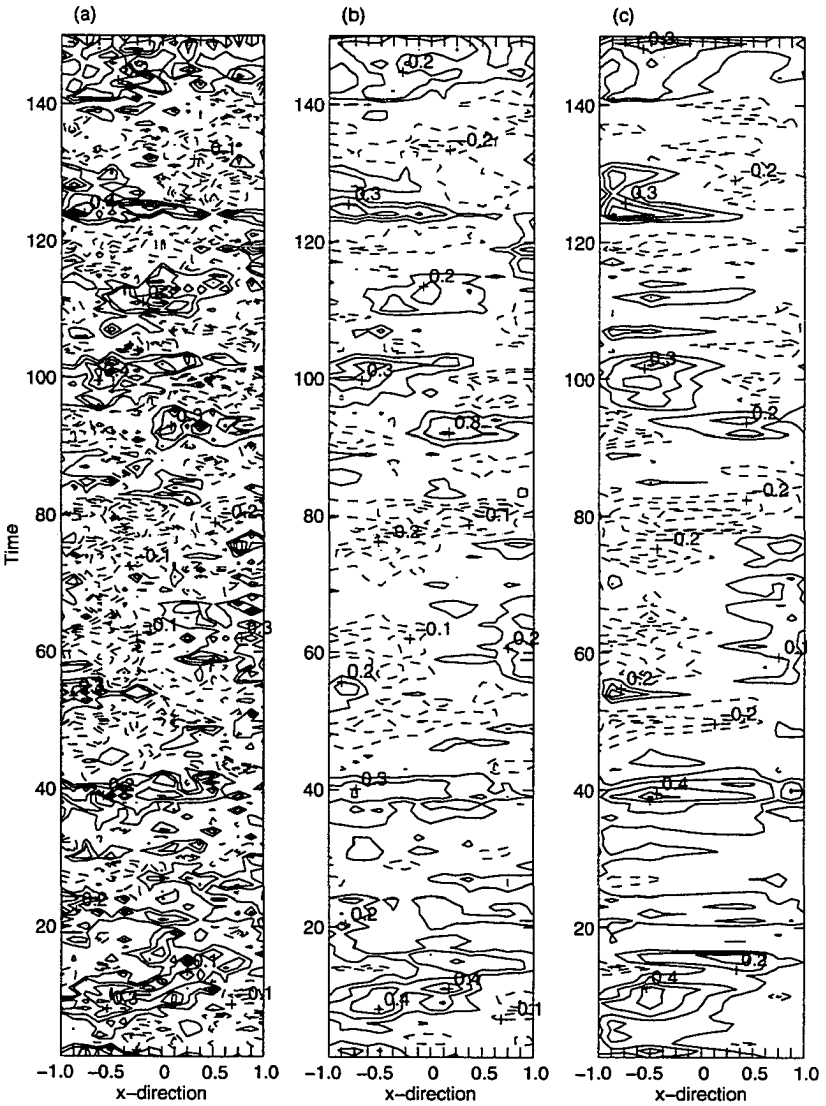

Figure 5 Simulation x-t contour plots at location $y = 0$ for the (a) noisy process Z,(b) true signal Y, and (c) prediction $\hat{Y}$ of the true signal given the noisy observations. Positive contours are indicated by the solid lines and negative contours are indicated by the dashed lines. The contour interval is 0.1.

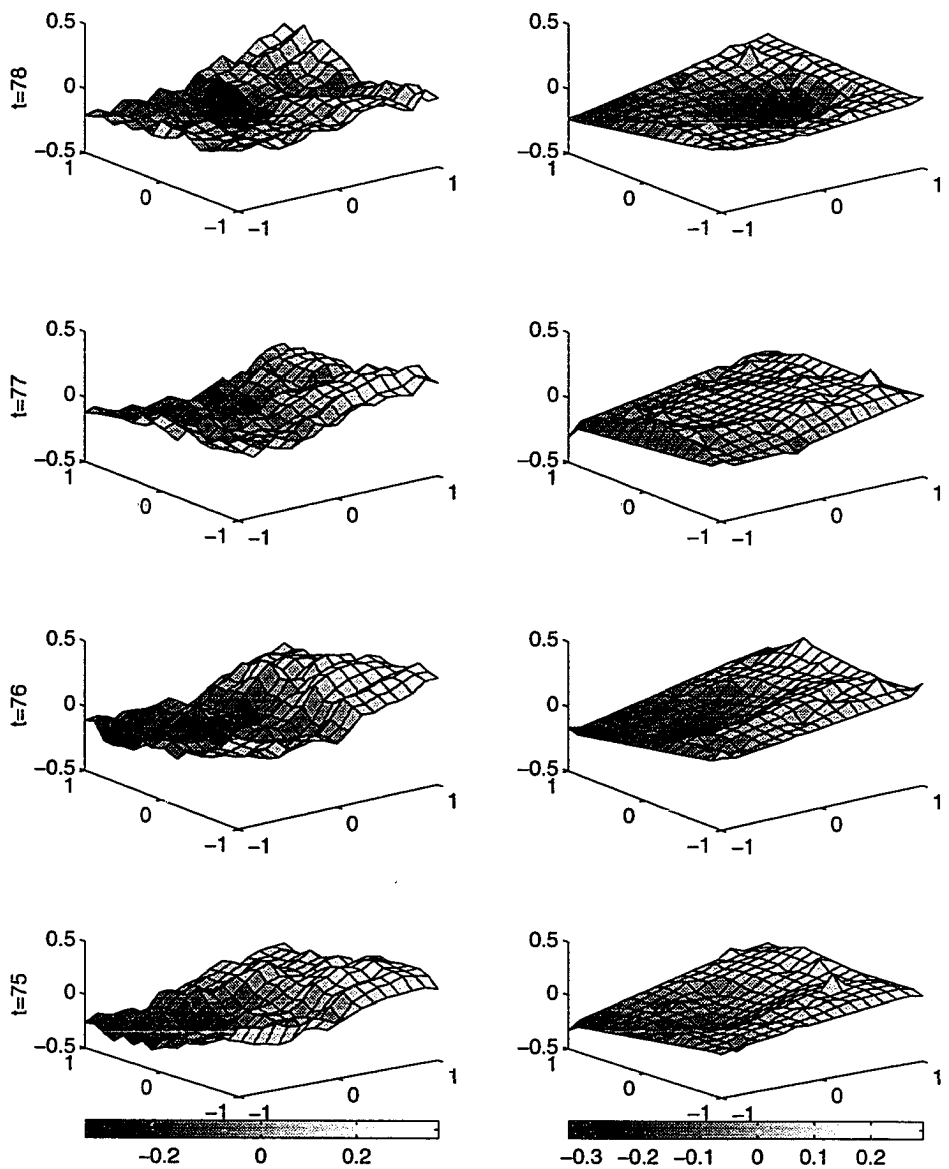

Figure 6 For the simulated data, the left side shows 3-d surface plots of the true signal $Y(\cdot, t)$ at times $t = 75, 76, 77, 78$; the right side shows the corresponding prediction $\hat{Y}(\cdot, t)$ of the true signal.

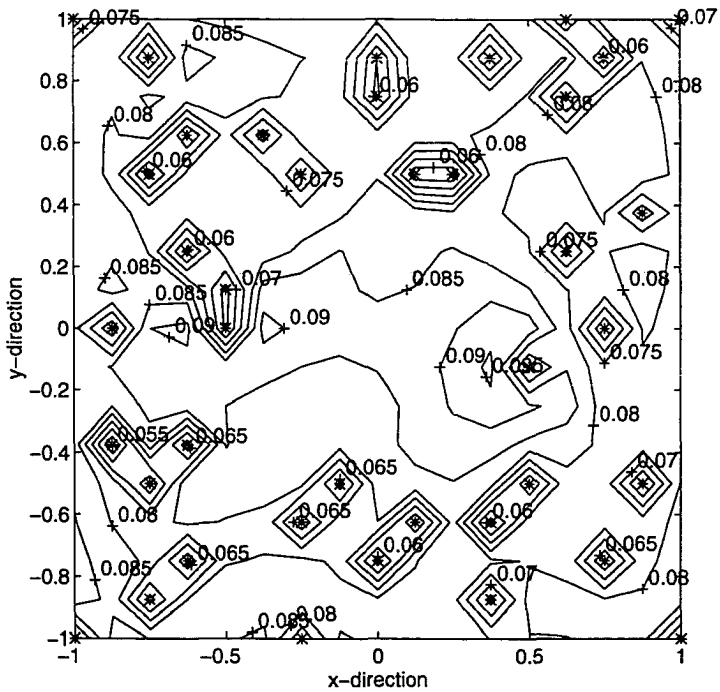

Figure 7 The SDTD mean squared prediction error standard deviations for the simulated data set. The contour interval is 0.005.

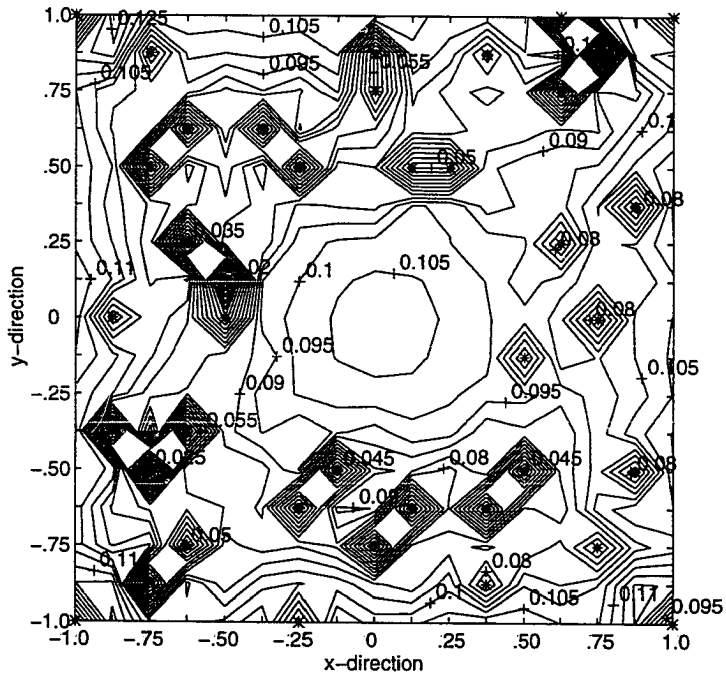

Figure 8 The simple kriging mean squared prediction error standard deviations for the simulated data set. The contour interval is 0.005.

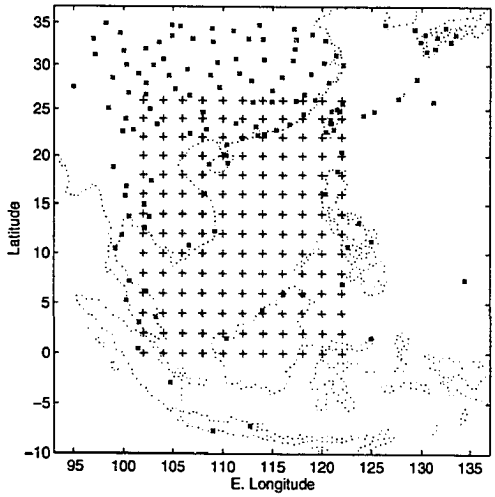

Figure 9 Observation locations (∗) and prediction grid (+) for the South China Sea precipitation example.

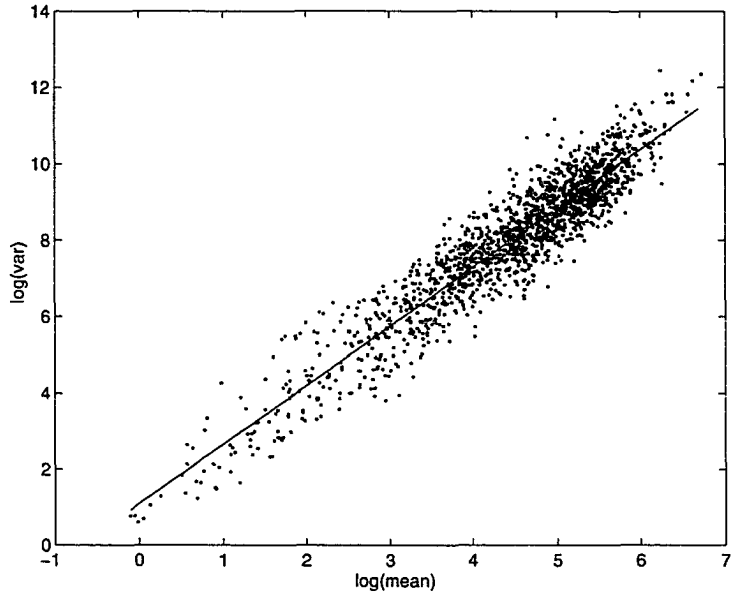

Figure 10 Natural log of precipitation (mm) mean vs. the natural log of the precipitation variance and the weighted least squares fit. Means and variances are calculated over time at each observation location.

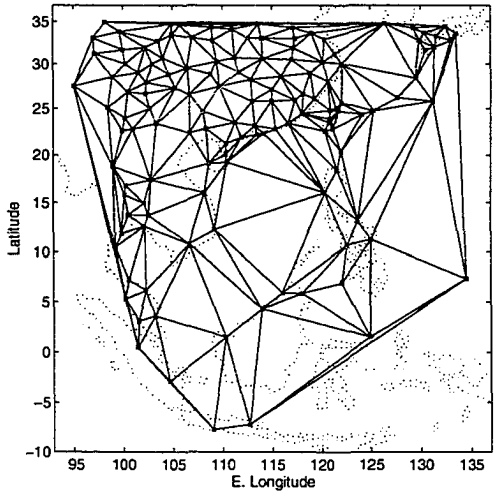

Figure 11 Delaunay triangulation of precipitation observation locations on a cylindrical equal area map projection.

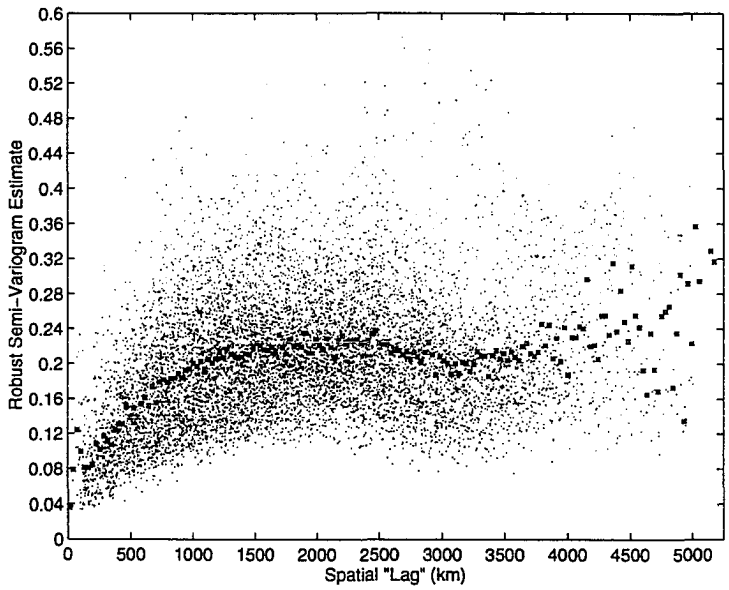

Figure 12 Robust semi-variogram estimate of the transformed precipitation data (·) and "bin" averages (∗). Each dot represents the average of variogram estimates over time.

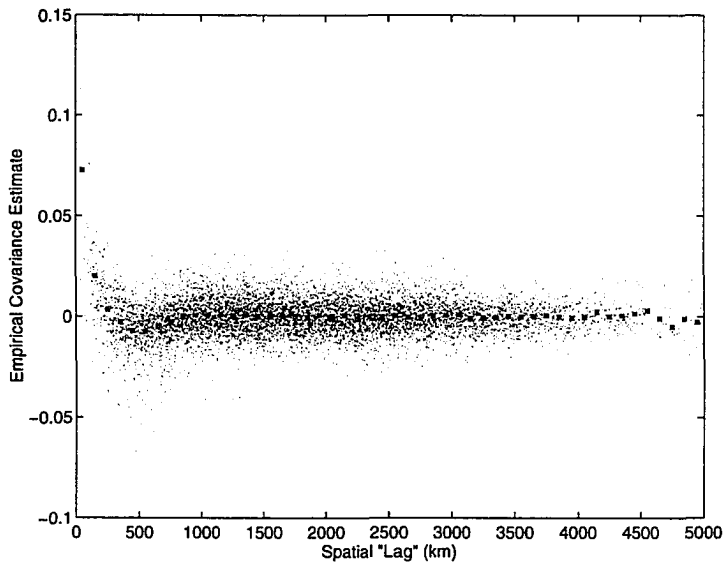

Figure 13 Estimated covariances of the ν process for the transformed precipitation process (·) and "bin" averages (∗).

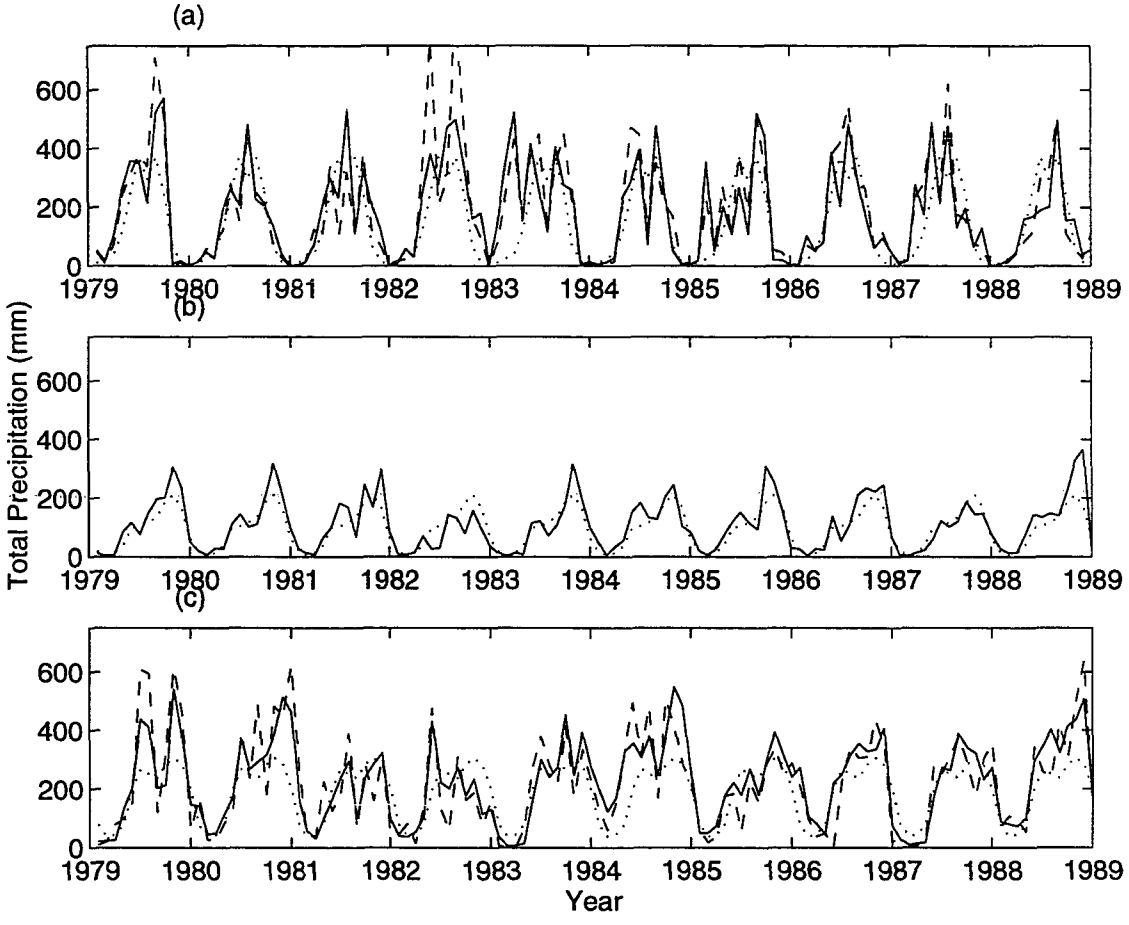

Figure 14 Time series plots for three locations over the 10-year period from January 1979 to December 1988: (a) a location on the southern coast of China near (114°E,22°N), (b) a location in the middle of the South China Sea (114°E,12°N), and (c) a location along the northwestern coast of Borneo near (114°E,4°N). The predicted precipitation (mm) is indicated by a solid line, the noisy observation (mm) [for (a) and (c)] by a dashed line, and the estimated seasonal mean precipitation (mm) by the dotted line.

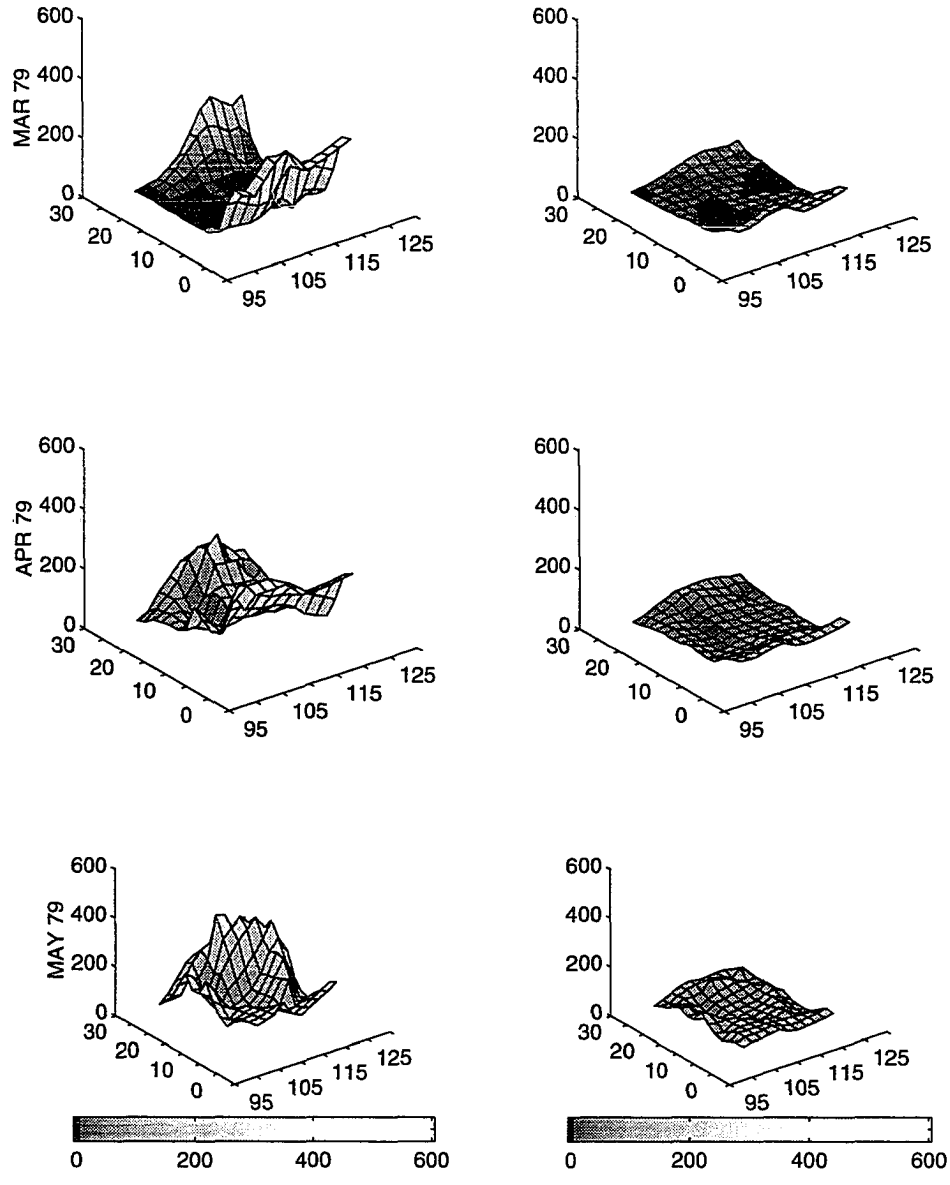

Figure 15 The left side shows 3-d surface plots of the precipitation prediction (mm) for March, April, and May of 1979; the right side shows the corresponding mean squared prediction error standard deviations.

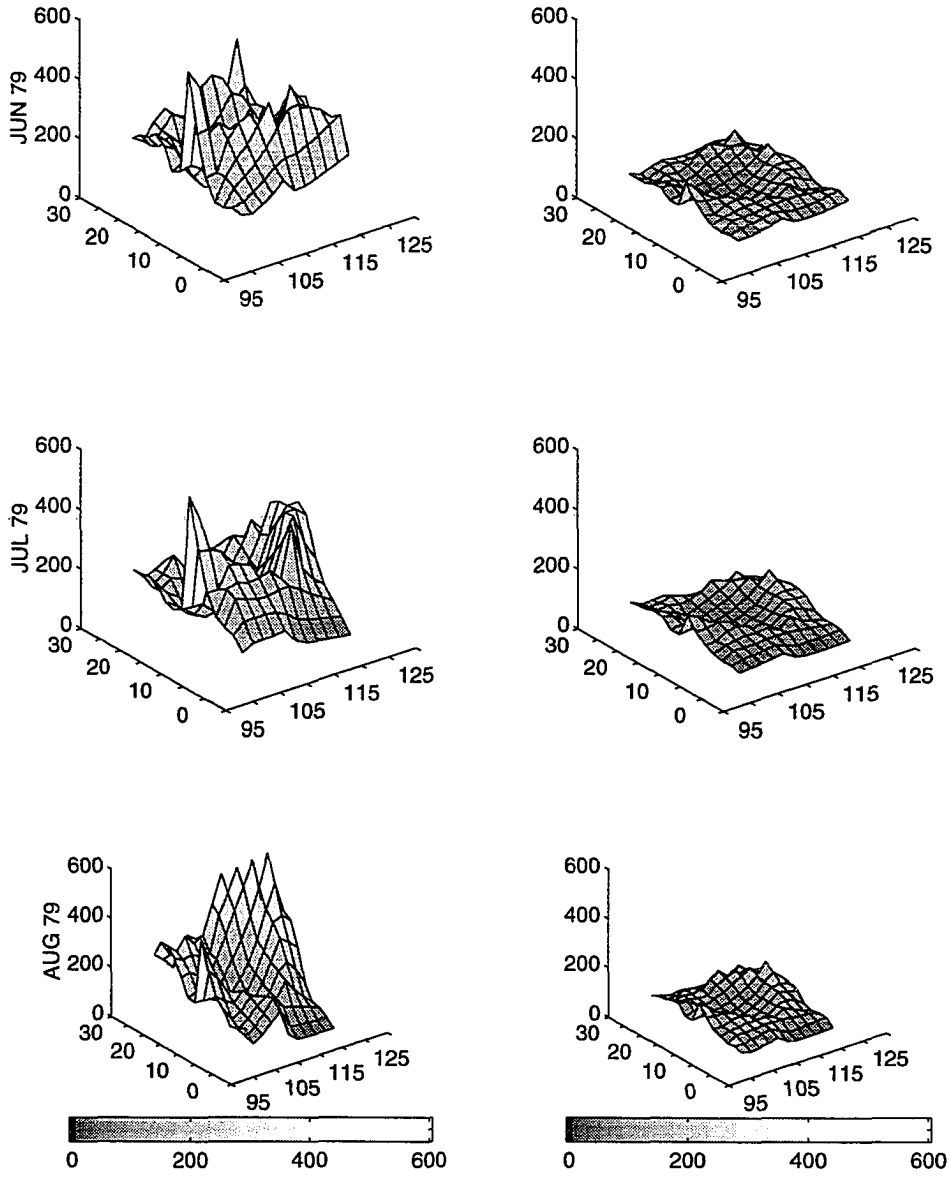

Figure 16 The left side shows 3-d surface plots of the precipitation prediction (mm) for June, July, and August of 1979; the right side shows the corresponding mean squared prediction error standard deviations.

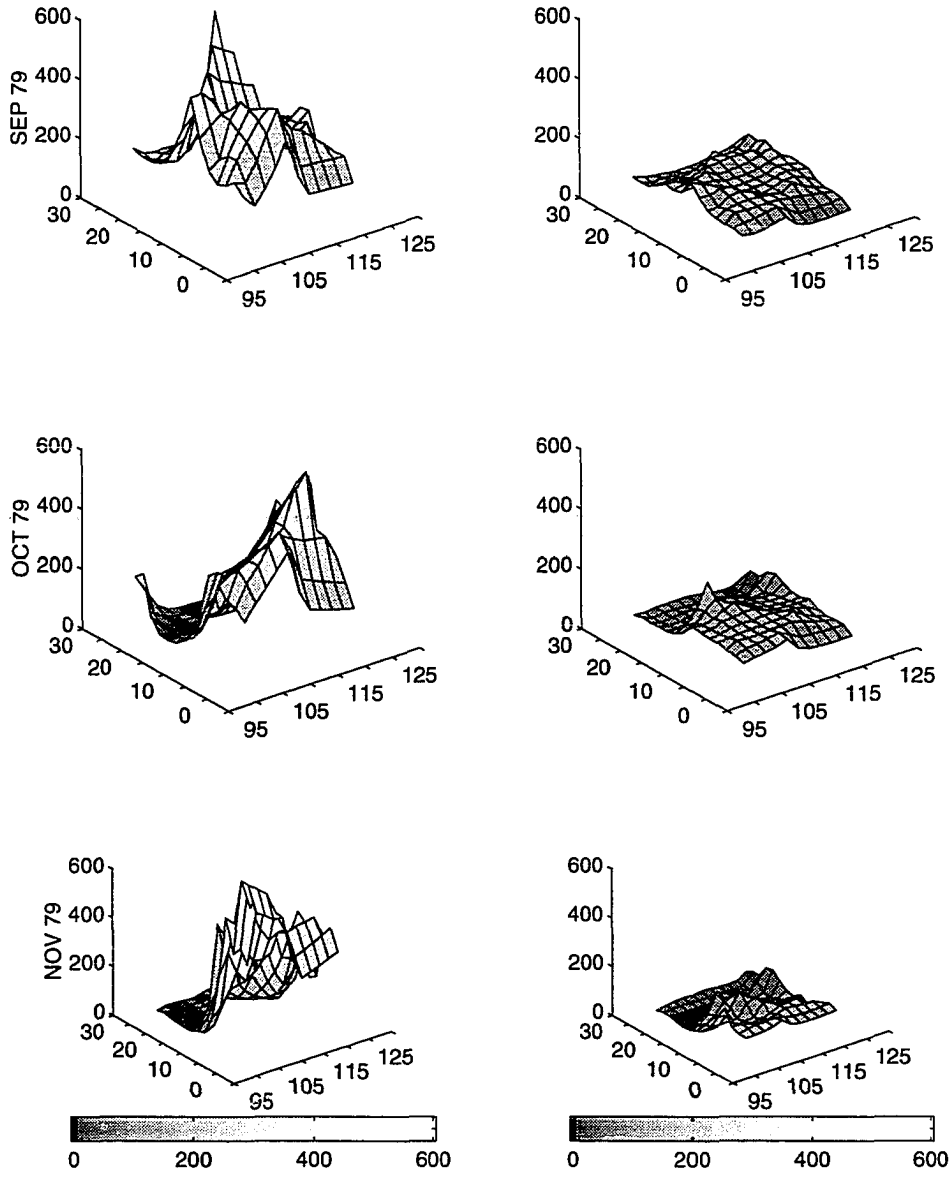

Figure 17 The left side shows 3-d surface plots of the precipitation prediction (mm) for September, October, and November of 1979; the right side shows the corresponding mean squared prediction error standard deviations.

GENERAL CONCLUSION

Meteorological and climatological processes typically show variability over both space and time. Consequently, scientists who study these processes require spatio-temporal models that can characterize this variability. Such characterizations then help the scientists to make inference about the nature of the processes, and eventually aid in their prediction. Thus, my goal in this dissertation has been to help bridge the gap between the need for advanced spatio-temporal statistical methods in the atmospheric sciences, and the development of such methods in the statistical sciences.

The dissertation consists of a background chapter followed by three additional chapters, each of which represents an independent paper. The background chapter is an overview of spatio-temporal statistical methods commonly used in the atmospheric sciences, from a statistical point of view. The first paper uses some simple harmonic analysis ideas to make spatial inference about a possible physical mechanism for the observed semiannual oscillation in the Northern Hemisphere extratropical height field. The second paper uses some advanced cyclostationary time series techniques to explore the seasonal variability of lower stratospheric mixed Rossby-gravity waves over the tropical Pacific. Finally, the third paper presents a new spatio-temporal statistical model that can be used to predict in both space and time. The results from each of these studies are described briefly below.

In the first chapter, a comprehensive overview of spatio-temporal methods that have received attention in the atmospheric science literature is presented. In particular, the focus is on the Empirical Orthogonal Function (EOF), Principal Interaction Pattern (PIP), Principal Oscillation Pattern (POP), and spatio-temporal Canonical Correlation Analysis (CCA) methods.

Particular attention is given to the physical and statistical considerations that must be made when applying these techniques to data. Specifically, these include the consideration of discrete or continuous space, the inclusion of measurement error, whether the application is prognostic or diagnostic, and whether the goal is spatial prediction, temporal prediction, or smoothing. We also consider what it is that is being optimized in a particular method, whether a dynamical component is appropriate, and whether Bayesian (i.e., Kalman filter) ideas should be considered. In general, most of these issues must be considered simultaneously. Unfortunately, this is rarely done in practice. Throughout the paper, we present several potential questions that are deserving of additional research. These generally are concerned with continuous space applications, and with the inclusion of measurement error in the analysis.

In the second chapter (i.e., the first paper) simple harmonic analysis is used to make diagnostic inference about the spatial variation of the semiannual component of the atmospheric general circulation. In particular, based on an examination of the spatial distribution of the maximum semiannual oscillation (SAO) amplitudes in the Northern Hemisphere (NH) extratropical 500-hPa height field, we concluded that this oscillation exhibits a very dominant zonally asymmetric east-west spatial structure. This led us to conclude that the stationary eddies inherent in the NH general circulation might be a useful tool with which to explain the SAO. Indeed, a comparison between the stationary eddies and the SAO showed that the NH midlatitude SAO can be explained almost entirely by the spatial and temporal asymmetries in the annual variation of the stationary eddies. It was suggested that the mechanism for the SAO in the NH extratropics is simply a result of land-sea contrasts, similar to the well-known explanation of the Southern Hemisphere (SH) SAO. However, the NH SAO is likely due to east-west contrasts between the continental land masses and oceans in the NH, while the SH SAO is due to the north-south contrast between the Antarctica land mass and the surrounding ocean. Thus, we have unified the conceptual view of the atmospheric SAO in both hemispheres. It was suggested that additional research should be focused on modeling studies with which these hypotheses could be verified, as well as additional focus on the NH polar SAO.

The third chapter (i.e., the second paper) is concerned with the seasonal variability of mixed Rossby-gravity waves (MRGWs) in the lower stratosphere over the tropical western Pacific. The study of these waves is important because they are believed to be a critical forcing mechanism of the quasi-biennial oscillation (QBO) in the tropical stratosphere. Recent observational studies have suggested that MRGWs generally do not show semiannual variability, but rather have a single seasonal peak, depending on geographic location. However, since these waves are believed to be associated with tropical convective activity, and since this convective activity exhibits semiannual peaks, we hypothesized that MRGWs should show semiannual peaks. Thus, we employed some relatively sophisticated time series analysis techniques to study the seasonal variability of these waves. These analyses used long time records of wind data in the lower stratosphere at four tropical Pacific observation stations. Specifically, seasonally varying cross-spectral analysis suggested that there are significant twice-yearly peaks in the v-wind power and the mean squared coherence between the u- and v-winds, with peaks occurring in the winter-early spring and in summer-early fall. In addition, the seasonally varying phase associated with the mean squared coherence analysis suggested that there is convergence of horizontal momentum flux associated with these waves, and that the sign of the convergence is opposite during the two seasonal maxima. This convergence of momentum flux in MRGWs has not been identified previously and deserves additional research effort. Furthermore, an autoregressive cyclic spectral estimate showed that the frequency of the maximum v-wind power in the MRGW frequency band shifts seasonally. This shift may be related to the seasonal variation of MRGWs, although further effort is required to prove such a claim.

In the fourth chapter (i.e., the third paper), a new spatio-temporal statistical model is proposed that attempts to consider the influence of both temporal and spatial variability. This model is mainly concerned with prediction, unlike the traditional spatio-temporal methods used in the atmospheric science literature (and outlined in the first chapter of this dissertation), which were primarily designed for diagnostic applications. The model is developed within the framework of continuous space and discrete time. The model then assumes a first-order

Markov temporal dynamic structure in conjunction with a spatially descriptive colored noise process. With the inclusion of a measurement error equation, we developed a spatio-temporal Kalman filter prediction algorithm that allowed us to predict in time and at spatial locations at which we do not have observations. The model prediction equation is quite general and includes a simple kriging analog as a special case. The model was shown to predict well with a simulated spatio-temporal data set, and was shown to be superior to simple kriging applied independently at each time. The model was then applied to a large precipitation data set in the South China Sea region. Predictions of precipitation over the data-sparse South China Sea seemed to capture the dynamic variation of the spatial precipitation field. Since this work is quite new, there are many possible avenues for future research. These were outlined in the paper.

ACKNOWLEDGEMENTS

I would never have been able to accomplish this task without the understanding and support of my family. First and foremost, I would like to express my thanks to Carolyn, the value of whose love and support throughout the past five years has been incalculable. I am forever grateful for her sacrifice as we placed much of our lives on hold through this period so that I could pursue a dream. I also thank my two wonderful children, Olivia and Nathan, who have never known a different life. I have appreciated their understanding and patience throughout this process and for their always being there to brighten my day.

I am eternally grateful to my parents, Bayliss and Irene Wikle, for instilling in me the value of education, and for giving me the freedom and support to search for my own path. Their support and love has shown me what it means to be a parent. In addition, I would like to thank my in-laws for their support, encouragement, and understanding over the years. I am exceedingly fortunate to have *two* such wonderful families.

I wish to thank by co-major professors, T.-C. (Mike) Chen and Noel Cressie. In addition to being top scientists in their respective fields, both have mastered the art of teaching, and the more elusive skill of mentoring. In countless informal conversations, Dr. Chen has taught me more than just the art of diagnostic analysis in atmospheric science, he has taught me the *philosophy of science* and the importance of knowing the history of your field. His support and encouragement in good times and in bad have been a source of strength. Dr. Cressie has provided countless points of wisdom and encouragement over the last couple of years, as well. In addition to opening my eyes to the wonders of spatial analysis, he has taught me how to think like a statistician and has provided a model of professionalism and scientific optimism

that I shall strive to emulate throughout my career. I especially appreciate the extra time, effort, and encouragement he has given over the last weeks during the completion of my degree.

For taking the time out of their busy schedules, I would also like to thank my committee members: Dr. Roland Madden of the National Center for Atmospheric Research, Dr. Peter Sherman, Dr. Raymond Arritt, and Dr. William Gutowski. In particular, Dr. Sherman has spent a grat deal of time teaching me about modern digital signal processing, and provided encouragement throughout the growing pains of my first scientific paper. I also would like to thank Dr. Madden for generously giving his time to teach me how to analyze physical time series and for setting the standard for using such analyses to answer questions about the physical world.

Of course, much of what has made the last five years such a positive experience was the friendship and support of my fellow students. In particular, I would like to thank Jean Pelkey, Alkhalil Adoum, Craig Clark, Zaitao Pan, Susan Kiehne, and John Iselin.

Finally, all of this would not have been possible without the generous financial support of the Graduate Fellowships for Global Change Program of the U.S. Department of Energy, as administered by Oak Ridge Institute for Science and Education. Rather than just talk about the need for interdisciplinary training in the sciences, they have funded it! I hope that I can prove that their substantial investment has not been in vain.